AD________________

Award Number: DAMD17-00-1-0197

TITLE: Computer-Aided Diagnosis of Breast Lesions

PRINCIPAL INVESTIGATOR: Yulei Jiang, Ph.D.

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, Illinois 60637

REPORT DATE: June 2002

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** June 2002	**3. REPORT TYPE AND DATES COVERED** Annual Summary (1 Jun 01 -31 May 02)

4. TITLE AND SUBTITLE
Computer-Aided Diagnosis of Breast Lesions

5. FUNDING NUMBERS
DAMD17-00-1-0197

6. AUTHOR(S):
Yulei Jiang, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Chicago
Chicago, Illinois 60637

E-Mail: y-jiang@uchicago.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

none provided

14. SUBJECT TERMS
breast lesions

15. NUMBER OF PAGES
48

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	Unlimited

TABLE OF CONTENTS

INTRODUCTION

The long-term goal of our research is to develop computer-aided diagnosis (CAD) techniques to improve the detection and diagnosis of breast cancer. The hypothesis to be tested in the present project is that radiologists' ability to differentiate malignant from benign breast lesions can be improved by integrating radiologists' perceptual expertise in the interpretation of mammograms with the advantages of automated computer classification. This project has 3 objectives:

1. To combine radiologist-extracted Breast Imaging Reporting and Data System (BI-RADS) features with image features extracted by a computer to classify malignant and benign clustered microcalcifications in mammograms.
2. To optimally combine radiologists' diagnosis with the result of computer classification.
3. To optimize computer classification for full-field digital mammograms.

BODY

1. Analysis of clinical benefits of CAD

We analyzed data obtained in a previous observer study to find potential clinical benefits from CAD in addition to what has already been demonstrated [1]. In this observer study, 10 radiologists reviewed mammograms of 104 patients both without and with a computer aid, which was designed to help radiologists differentiate malignant from benign clustered microcalcifications in mammograms [2]. Previously, we demonstrated that this computer aid helped the radiologists to improve diagnostic accuracy. Specifically, the computer aid helped each of the radiologists to recommend, on average, 14% more biopsies for malignant lesions and 10% fewer biopsies for benign lesions [1]. In the present analysis, we demonstrated that this computer aid helped radiologists substantially reduce their variability in the interpretation of mammograms. Because of the lack of a universal metric to quantify the degree of variability in the interpretation of mammograms, we used two different approaches in this analysis. An empirical analysis is published in Radiology (reprint included with report) [3]. Part of this work was also presented at an annual meeting of the American Association of Physicists in Medicine (AAPM) [4] and the 86th Scientific Assembly and Annual Meeting of Radiological Society of North America

(RSNA) [5]. In addition, Dr. Robert F. Wagner, who with colleagues has developed a component-of-variance model that can be estimated from observer study data, analyzed our observer study data and found the same conclusion that CAD reduced reader variability in our study (reprint included with report) [6]. Furthermore, we published a review article on the potential clinical benefits of breast cancer CAD (reprint included with report) [7].

2. Comparison of BI-RADS lesion descriptors and computer-extracted image features

We have been conducting an ongoing study to compare computer-extracted image features that we have developed previously [1, 2, 8] and BI-RADS lesion descriptors [9, 10] for computer classification of breast lesions as malignant or benign. Our goal was to identify the relative strengths and weaknesses of these two sources of image features for computer classification and potentially improve the performance of computer classification by combining the image features from both sources. We investigate both clustered microcalcifications and masses in this study, even though we proposed in the original grant application to study only clustered microcalcifications.

We have collected a database (additional images are not yet analyzed) of 67 mammograms containing masses (33 malignant) and 99 mammograms containing microcalcifications (42 malignant), with each case composed of original mammograms in the standard and magnification or spot-compression views. Two expert mammographers who are familiar with BI-RADS have reviewed these cases and provided their descriptions of the lesions and their final assessments in terms of BI-RADS. Analysis showed that: (1) the radiologists were more accurate at diagnosing masses than at diagnosing clustered microcalcifications, (2) computer classification of breast lesions as malignant or benign for masses tended to be more accurate based on BI-RADS lesion descriptions provided by radiologists, and for microcalcifications tended to be more accurate based on computer-extracted image features, and (3) computer classification achieved the best performance based on the combination of BI-RADS lesion descriptions provided by radiologists and computer-extracted image features. Part of this work has been presented or will be presented at the 6th International Workshop on Digital Mammography in Bremen, Germany (preprint included with report) [11, 12], the Department of Defense Era of Hope Breast Cancer Research Program Meeting [13, 14], and the 88th Scientific Assembly and Annual Meeting of the RSNA [15].

We have also been developing a method to use a computer to predict BI-RADS lesion descriptions that a radiologist would most likely use to describe microcalcifications in a mammogram and have achieved limited success. In this method, we selected a set of computer-extracted image features and used linear discriminante analysis (LDA) classifiers to assign a "score" to each BI-RADS lesion description terms. We then selected the one or two BI-RADS terms that were assigned the largest "scores" as the computer prediction of what a radiologist would most likely use to describe the microcalcifications. Our analysis showed that the concordance between the computer prediction and data obtained from radiologists is about the same as the concordance between two radiologists. Part of this work was presented at the 44th Annual Meeting of the AAPM [16].

3. "Optimal" combination of radiologists' and a computer's diagnostic assessment

We have been developing a method to "optimally" combine the quantitative diagnostic assessments made by a radiologist and by a computer, based on a bivariate binormal model that was originally developed for ROC analysis [17]. This method takes into account the individual accuracy of the radiologist and the computer, as well as the correlation between their diagnostic assessments. Previously, we evaluated this method on a dataset obtained from a mammography observer study and found that the method produced better results than what radiologists have achieved by using the computer aid in an *ad hoc* way [18, 19]. We have now evaluated this method on a second dataset from a chest radiograph observer study and have found a similar conclusion that this method produced better results than both radiologists and the computer aid and, therefore, potentially better than what radiologists can achieve by using the computer aid in an *ad hoc* way. Part of this work was presented at the 87th Scientific Assembly and Annual Meeting of the RSNA [20].

4. Variability of the outputs of artificial neural network

While not proposed in our original grant application, we have started a study of the fundamental properties of artificial neural networks (ANNs) because ANNs are frequently used in CAD techniques as classifiers and they are used in the techniques that we have developed to classify breast lesions as malignant or benign [1, 2, 8]. We recently discovered that there is variability in the outputs from ANNs in an exact analogy as the uncertainties associated invariably with statistical estimates. This fundamental property of the ANNs either has not been recognized before or has been ignored in medical imaging research, where the main focus has been on the overall accuracy of a classifier. We have found that while one can train multiple ANNs based on a given training dataset to achieve highly similar overall performance (e.g., as measured by the A_z value), the outputs from these ANNs tend to be much more variable (by about two orders of magnitude). Therefore, such variability would logically have practical implications, and likely detrimental effects, on the interpretation of the ANN outputs, and ultimate may affect breast cancer diagnosis when CAD is used. This aspect of the ANNs should, therefore, be a consideration in designing CAD techniques. This work was presented at the Medical Image Perception Conference IX [21] and at the 44th Annual Meeting of the AAPM [22]. We have submitted a manuscript to *IEEE Transaction on Medical Imaging* [23], which will be included with our next report once it is published.

5. An empirical comparison of Student's t-test and the Dorfman-Berbaum-Metz method

We have conducted another study not originally proposed in our grant application, to compare the Student's t-test and the Dorfman-Berbaum-Metz (DBM) method for comparing competing diagnostic modalities. Both methods are frequently used in the evaluation of CAD in observer studies. Theoretically, the t-test is not appropriate for such comparisons because it takes into account only the reader variance and ignores the case variance [24], whereas the DBM (and other similar methods) are more appropriate because it takes into account both the reader and the case variance [25]. Therefore, strictly speaking, a statistical conclusion from the t-test can be generalized to a population of readers being studied but only for the specific cases being studied, whereas a statistical conclusion from the DBM method can be generalized to a population of readers and a population of cases being studied, which is generally the objective of observer studies.

We analyzed three CAD observer study datasets and compared the results of statistical analysis from the t-test and from the DBM method by sampling readers and cases from the larger pool of data, resulting in several million individual comparisons. We found that the results from the t-test and from the DBM method are often similar; however, they can frequently be different by a larger amount. Therefore, it is indeed not appropriate to use the t-test on both theoretical grounds and on the grounds of our empirical analysis. This work was presented at the SPIE's International Symposium: Medical

Imaging 2002 (reprint included with report) [26]. We are preparing a manuscript to be submitted to *Academic Radiology*, and will include the manuscript in our next report once it is published.

KEY RESEARCH ACCOMPLISHMENTS

- Analysis of potential clinical benefits of CAD of malignant and benign breast lesions.
- Initial comparison of BI-RADS lesion descriptors provided by radiologists and computer-extracted image features for computer classification of breast lesions as malignant or benign.
- Development of a method to predict BI-RADS lesion descriptors that a radiologist would most likely use to describe microcalcifications in a mammogram.
- Development of a novel method for the "optimal" combination of quantitative diagnostic assessments made by a radiologist and made by a computer.
- Investigation of the variability in the outputs of artificial neural networks used in breast cancer CAD.
- An empirical comparison of the Student's t-test and the Dorfman-Berbaum-Metz method for statistical comparison of competing diagnostic modalities involving CAD.

REPORTABLE OUTCOMES

1. Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. The potential of computer-aided diagnosis (CAD) to reduce variability in radiologists' interpretation of mammograms. *Radiology* 220:787-794, 2001.
2. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities. Acad Radiol 8:605-615, 2001.
3. Jiang Y. Computer-aided diagnosis of breast cancer in mammography: evidence and potential. *Technology in Cancer Research and Treatment* 1:211-216, 2002.
4. Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ, Giger ML, Lan L, Huo Z, Edwards AV. Comparison of BI-RADS lesion descriptors and computer-extracted image features for computer classification of malignant and benign breast lesions. Presented at the *6th International workshop on Digital Mammography*, Bremen, Germany, 2002.

5. Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ, Giger ML, Lan L, Huo Z, Edwards AV. Comparison of BI-RADS lesion descriptors and computer-extracted image features for computer classification of malignant and benign breast lesions. In Peitgen HO, Ed., *Digital Mammography 2002* Heidelberg: Springer Verlag Publishers, 2002.
6. Jiang Y, Paquerault S, Nishikawa RM, Giger ML, Schmidt RA, D'Orsi CJ, Vyborny CJ, Metz CE. Computer-aided diagnosis of malignant and benign breast lesions in mammograms. Invited symposium platform presentation at the *Era of Hope 2002 Department of Defense Breast Cancer Research Program Meeting*, Orlando, FL, 2002.
7. Jiang Y, Paquerault S, Nishikawa RM, Giger ML, Schmidt RA, D'Orsi CJ, Vyborny CJ, Metz CE. Computer-aided diagnosis of malignant and benign breast lesions in mammograms. Poster presented at the *Era of Hope 2002 Department of Defense Breast Cancer Research Program Meeting*, Orlando, FL, 2002.
8. Jiang Y, Schmidt RA, D'Orsi CJ, Vyborny CJ, Nishikawa RM, Paquerault S. Classification of malignant and benign clustered microcalcifications based on computer-extracted lesion features and radiologist-provided BI-RADS description. *Radiology* 225(P):, 2002. Presented at the *88th Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, 2002.
9. Paquerault S, Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ. Computer BI-RADS analysis of clustered microcalcifications in mammograms. *Medical Physics* 29:, 2002.
10. Jiang Y, Metz CE. A new method for combining radiologists' and a computer's diagnostic assessments. *Radiology* 221(P):424, 2001. Presented at the *87th Scientific Assembly and Annual Meeting of Radiological Society of North America*, Chicago, IL, 2001.
11. Jiang Y. Uncertainty in the output of artificial neural networks. Presented at *Medical Image Perception Conference IX*, Airlie Conference Center, Warrenton, VA, 2001.
12. Jiang Y. Uncertainty of artificial neural network output. *Medical Physics* 29:1323, 2002. Presented at the *44th Annual Meeting of the American Association of Physicists in Medicine*, Montreal, Canada, 2002.

13. Jiang Y. Uncertainty in the output of artificial neural networks. *IEEE Transactions on Medical Imaging* (submitted August 2002).
14. Jiang Y. Comparison of student's t-test and the Dorfman-Berbaum-Metz (DBM) method for the statistical comparison of competing diagnostic modalities. *Proc. SPIE* 4686:205-209, 2002.

CONCLUSIONS

We have made progress toward the objectives of this project. The research results are positive and support project continuation.

REFERENCES

1. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 6:22-33, 1999.
2. Jiang Y, Nishikawa RM, Wolverton DE, Metz CE, Giger ML, Schmidt RA, Vyborny CJ, Doi K. Malignant and benign clustered microcalcifications: Automated feature analysis and classification. *Radiology* 198:671-678, 1996.
3. Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. The potential of computer-aided diagnosis (CAD) to reduce variability in radiologists' interpretation of mammograms. *Radiology* 220:787-794, 2001.
4. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Doi K. Multiple benefits of computer-aided diagnosis (CAD) in the diagnosis of malignant and benign breast lesions. Presented at World Congress on Medical Physics and Biomedical Engineering, July 2000.
5. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Doi K. Three potential benefits of computer-aided diagnosis (CAD) in breast cancer diagnosis. Chicago, Illinois: the 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, 2000.
6. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities. Acad Radiol 8:605-615, 2001.

7. Jiang Y. Computer-aided diagnosis of breast cancer in mammography: evidence and potential. *Technology in Cancer Research and Treatment* 1:211-216, 2002.

8. Huo Z, Giger ML, Vyborny CJ, Metz CE. Breast Cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms. Radiology 224:560-568, 2002.

9. American College of Radiology (ACR). Breast imaging reporting and data system (BI-RADSTM). Vol. Third Edition ed. Reston, VA: American College of Radiology, pp. 1998.

10. Baker JA, Kornguth PJ, Lo JY, Floyd CEJ. Artificial neural network: improving the quality of breast biopsy recommendations. Radiology 198:131-135, 1996.

11. Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ, Giger ML, Lan L, Huo Z, Edwards AV. Comparison of BI-RADS lesion descriptors and computer-extracted image features for computer classification of malignant and benign breast lesions. Presented at the *6th International workshop on Digital Mammography*, Bremen, Germany, 2002.

12. Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ, Giger ML, Lan L, Huo Z, Edwards AV. Comparison of BI-RADS lesion descriptors and computer-extracted image features for computer classification of malignant and benign breast lesions. In Peitgen HO, Ed., *Digital Mammography 2002* Heidelberg: Springer Verlag Publishers, 2002.

13. Jiang Y, Paquerault S, Nishikawa RM, Giger ML, Schmidt RA, D'Orsi CJ, Vyborny CJ, Metz CE. Computer-aided diagnosis of malignant and benign breast lesions in mammograms. Invited symposium platform presentation at the *Era of Hope 2002 Department of Defense Breast Cancer Research Program Meeting*, Orlando, FL, 2002.

14. Jiang Y, Paquerault S, Nishikawa RM, Giger ML, Schmidt RA, D'Orsi CJ, Vyborny CJ, Metz CE. Computer-aided diagnosis of malignant and benign breast lesions in mammograms. Poster presented at the *Era of Hope 2002 Department of Defense Breast Cancer Research Program Meeting*, Orlando, FL, 2002.

15. Jiang Y, Schmidt RA, D'Orsi CJ, Vyborny CJ, Nishikawa RM, Paquerault S. Classification of malignant and benign clustered microcalcifications based on computer-extracted lesion features and radiologist-provided BI-RADS description. *Radiology* 225(P):, 2002. Presented at the *88th Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, 2002.

16. Paquerault S, Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ. Computer BI-RADS analysis of clustered microcalcifications in mammograms. *Medical Physics* 29:, 2002.

17. Metz CE, Wang P-L and Kronman HB, "A new approach for testing the significance of differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging,* edited by F. Deconinck, pp. 432-445, Nijhoff, The Hague, 1984.

18. Jiang Y, Metz CE. An optimal method for combining two correlated diagnostic assessments with application to computer-aided diagnosis. Presented at SPIE's International Symposium: Medical Imaging 2001, February 2001.

19. Jiang Y, Metz CE. An optimal method for combining two correlated diagnostic assessments with application to computer-aided diagnosis. Proc. SPIE 4324:177-183, 2001.

20. Jiang Y, Metz CE. A new method for combining radiologists' and a computer's diagnostic assessments. *Radiology* 221(P):424, 2001. Presented at the *87th Scientific Assembly and Annual Meeting of Radiological Society of North America*, Chicago, IL, 2001.

21. Jiang Y. Uncertainty in the output of artificial neural networks. Presented at *Medical Image Perception Conference IX*, Airlie Conference Center, Warrenton, VA, 2001.

22. Jiang Y. Uncertainty of artificial neural network output. *Medical Physics* 29:1323, 2002. Presented at the *44th Annual Meeting of the American Association of Physicists in Medicine*, Montreal, Canada, 2002.

23. Jiang Y. Uncertainty in the output of artificial neural networks. *IEEE Transactions on Medical Imaging* (submitted August 2002).

24. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234-245.

25. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27:723-731.

26. Jiang Y. Comparison of student's t-test and the Dorfman-Berbaum-Metz (DBM) method for the statistical comparison of competing diagnostic modalities. *Proc. SPIE* 4686:205-209, 2002.

LIST OF ATTACHED REPRINTS

1. Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. The potential of computer-aided diagnosis (CAD) to reduce variability in radiologists' interpretation of mammograms. *Radiology* 220:787-794, 2001.

2. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities. Acad Radiol 8:605-615, 2001.

3. Jiang Y. Computer-aided diagnosis of breast cancer in mammography: evidence and potential. *Technology in Cancer Research and Treatment* 1:211-216, 2002.

4. Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ, Giger ML, Lan L, Huo Z, Edwards AV. Comparison of BI-RADS lesion descriptors and computer-extracted image features for computer classification of malignant and benign breast lesions. In Peitgen HO, Ed., *Digital Mammography 2002* Heidelberg: Springer Verlag Publishers, 2002.

5. Jiang Y. Comparison of student's t-test and the Dorfman-Berbaum-Metz (DBM) method for the statistical comparison of competing diagnostic modalities. *Proc. SPIE* 4686:205-209, 2002.

Yulei Jiang, PhD
Robert M. Nishikawa, PhD
Robert A. Schmidt, MD[2]
Alicia Y. Toledano, ScD[3]
Kunio Doi, PhD

Potential of Computer-aided Diagnosis to Reduce Variability in Radiologists' Interpretations of Mammograms Depicting Microcalcifications[1]

Index terms:
Breast neoplasms, calcification, 00.81
Breast neoplasms, diagnosis, 00.30
Computers, diagnostic aid
Diagnostic radiology, observer performance

Published online: August 21, 2001
10.1148/radiol.2200011257
Radiology 2001; 220:787–794

Abbreviations:
A_z = area under the ROC curve
CAD = computer-aided diagnosis
ROC = receiver operating characteristic

PURPOSE: To evaluate whether computer-aided diagnosis can reduce interobserver variability in the interpretation of mammograms.

MATERIALS AND METHODS: Ten radiologists interpreted mammograms showing clustered microcalcifications in 104 patients. Decisions for biopsy or follow-up were made with and without a computer aid, and these decisions were compared. The computer was used to estimate the likelihood that a microcalcification cluster was due to a malignancy. Variability in the radiologists' recommendations for biopsy versus follow-up was then analyzed.

RESULTS: Variation in the radiologists' accuracy, as measured with the SD of the area under the receiver operating characteristic curve, was reduced by 46% with computer aid. Access to the computer aid increased the agreement among all observers from 13% to 32% of the total cases ($P < .001$), while the κ value increased from 0.19 to 0.41 ($P < .05$). Use of computer aid eliminated two-thirds of the substantial disagreements in which two radiologists recommended biopsy and routine screening in the same patient ($P < .05$).

CONCLUSION: In addition to its demonstrated potential to improve diagnostic accuracy, computer-aided diagnosis has the potential to reduce the variability among radiologists in the interpretation of mammograms.

Multiple investigators (1–4) have shown that considerable variability exists among radiologists in the interpretation of mammograms. This variability affects the diagnostic accuracy of radiologists, as measured with receiver operating characteristic (ROC) analysis. Moreover, it directly affects their clinical decisions to recommend either biopsy or follow-up. Because such variability decreases the clinical effectiveness of breast cancer screening, it should be eliminated whenever possible. Some (5–7) have suggested that computer-aided diagnosis (CAD), in which a radiologist combines an independent analysis of mammograms performed by using a computer technique with his or her own reading, can potentially reduce interpretation variability. However, to our knowledge, this potential of CAD has not yet been demonstrated. We analyzed data obtained in an observer study (8) to compare variabilities in the interpretation of mammograms with and without use of a computer aid. Previously, we analyzed the data of that observer study and found that radiologists can improve their diagnostic performance by using a computer aid (8). The purpose of this study was to evaluate whether CAD can reduce interobserver variability among radiologists in the interpretation of mammograms.

MATERIALS AND METHODS

Case Materials

We obtained (from the University of Chicago Hospitals, Illinois) 104 mammograms of 46 consecutive malignant and 58 consecutive benign clustered microcalcifications that

[1] From the Kurt Rossmann Laboratories for Radiologic Image Research, Dept of Radiology (Y.J., R.M.N., R.A.S., K.D.), and the Depts of Anesthesia and Critical Care (A.Y.T.) and Health Studies (A.Y.T.), Univ of Chicago, 5841 S Maryland Ave, MC2026, Chicago, IL 60637. Received Jul 13, 2000; revision requested Aug 21; final revision received Jan 15, 2001; accepted Feb 15. Supported in part by NIH grant CA 60187. **Address correspondence to** Y.J. (e-mail: *y-jiang@uchicago.edu*).

This work was performed as part of the International Digital Mammography Development Group. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations.

Current addresses:
[2] Dept of Radiology, New York Univ Medical Ctr, NY.
[3] Ctr for Statistical Sciences, Brown Univ, Providence, RI.

Author contributions:
Guarantor of integrity of entire study, Y.J.; study concepts, Y.J., R.M.N., R.A.S., K.D.; study design, Y.J., R.M.N., R.A.S.; literature research, Y.J., A.Y.T.; experimental studies, Y.J.; data acquisition, Y.J.; data analysis/interpretation, Y.J., A.Y.T.; statistical analysis, A.Y.T.; manuscript preparation and editing, Y.J.; manuscript definition of intellectual content, revision/review, and final version approval, all authors.

were examined at biopsy. Our institutional review board approved a waiver for patient consent for this study because our study involved only retrospective review of existing mammograms. We included only cases of microcalcification because our computer aid was specifically designed to analyze this common type of mammographic lesion (work on computer analysis of breast masses is ongoing [9]) and because microcalcifications are often the only mammographic indication of breast cancer (10).

Of the malignant cases, 37 were ductal carcinoma in situ, and nine were invasive ductal carcinoma. Of the benign cases, two were lobular carcinoma in situ, four were atypical ductal hyperplasia, 16 were hyperplasia without atypia, seven were adenosis, six were fibroadenoma, 18 were fibrocystic change or fibrosis, and five were breast tissue without specific abnormality.

Consecutive cases were collected by using the following criteria: *(a)* A cluster of microcalcifications was the only suspect lesion, which led to the biopsy and for which the pathologic results were definitive; *(b)* original mammograms, including at least two standard views and one magnification view, were available; and *(c)* the technical quality of the mammograms was adequate for interpretation (8). To balance the number of malignant and benign cases and thereby increase statistical power, the malignant cases were collected, necessarily, from a longer period (11,12). These cases were clinically evaluated before the Breast Imaging Reporting and Data System was implemented; therefore, they were not assigned a Breast Imaging Reporting and Data System assessment category (8). Additional specific details regarding case selection are reported elsewhere (8).

Radiologist Observers

Ten radiologists, who had experience in mammography but who had not previously seen the study cases, interpreted the mammograms. Five observers were practicing radiologists from the Chicago metropolitan area, and five were senior radiology residents from our institution. For the attending radiologists, mammography accounted for an average of 30% of their clinical practice, and they were certified readers according to the Mammography Quality Standards Act. They had been reading mammograms for an average of 9 years (median, 6 years; range, 1–30 years), and they had read at least 1,000 mammograms in the preceding year. The residents had limited experience from training rotations of 1–2 months duration. Written informed consent, as approved by our institutional review board, was obtained from all observers after the nature of the experiment was fully explained. Data analysis was performed for three observer groups: all observers (n = 10), attending radiologists (n = 5), and residents (n = 5).

Computer Aid

The computer aid was an estimate of the likelihood (0%–100%) that a microcalcification cluster was due to a malignancy. An artificial neural network calculated the estimate on the basis of eight image features that were automatically extracted from standard-view screen-film mammograms (13). Mammograms were digitized with a 0.1-mm pixel size and a 12-bit gray scale by using a digitizer (Lumiscan 100; Lumisys, Sunnyvale, Calif). Locations of microcalcifications were manually identified on a computer monitor (8).

The observers were explicitly instructed to use the computer aid in their interpretation. They were told that the computer output had a sensitivity (defined as the fraction of cancers for which biopsy would have been recommended) of approximately 90% and positive predictive value (defined as the fraction of all cases for which biopsy would have been recommended that were cancers) of approximately 61% when a threshold of 30% was applied to the computer-estimated likelihood of malignancy. The performance estimates of the computer were obtained from the study cases. One interpretation of this instruction is that any observer could have achieved the same accuracy as the computer by recommending biopsy only when the computer reported a likelihood of malignancy of 30% or greater.

Data Acquisition

Each observer reviewed the cases twice: once with and once without the computer aid; each review was separated by an average of 30 days (range, 10–60 days). The following counterbalanced study design was used: Half of the mammograms were read without the computer aid in the first reading session and were read again with the computer aid in the second reading session; the other half of the mammograms were read first with the computer aid and then without the aid. The study design minimizes potential biases; it has been well documented (11,12) and has been described (8) in detail. The observers were asked to report *(a)* their level of confidence (on an analog scale of 0%–100%) that a lesion was malignant and *(b)* their clinical recommendation (Table 1).

TABLE 1
Clinical Recommendations Available to the Observers

Option	Recommendation
a	Surgical biopsy
b	Alternative tissue sampling*
c	Short-term follow-up
d	Routine follow-up

* Stereotactic or ultrasonography-guided core biopsy or fine-needle aspiration.

Data Analyses

We assessed interpretation variability by using three methods: *(a)* sensitivity, specificity, and ROC analysis; *(b)* analysis of interobserver agreement; and *(c)* analysis of substantial disagreements in clinical recommendations. Interobserver variability was assessed in these analyses. Intraobserver variability was not measured because no observer repeated mammographic interpretation either with or without the computer aid. Custom software was used to perform all calculations except calculations of κ values, which were determined by using other software (SPLUS; MathSoft, Seattle, Wash). The Student t and McNemar χ^2 tests were used to calculate P values, and the bootstrap method was used to estimate the 95% CIs in the statistical analyses.

Sensitivity was defined as the fraction of cancers for which surgical biopsy or alternative tissue sampling was recommended. Specificity was defined as the fraction of benign lesions for which short-term or routine follow-up was recommended. Because sensitivity and specificity incompletely describe accuracy and because they depend on how a radiologist selects a decision threshold to define positive diagnoses, we also performed an ROC analysis, which is the standard method for evaluating observer accuracy (6,14,15). We obtained ROC curves by fitting the binormal model to the confidence data, and we obtained summary ROC curves for the 10 observers as a group by averaging the slope and intercept parameters of the individual curves (14). The area under the ROC curve (A_z) was used as a summary index of accuracy. A_z can have values between

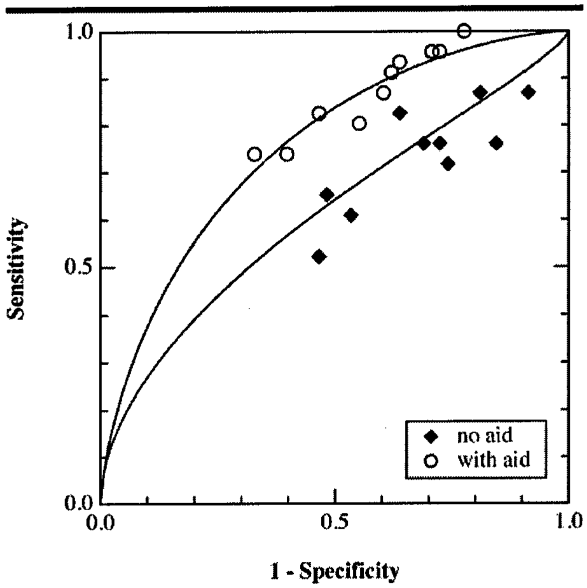

Figure 1. ROC curves and sensitivity and specificity data obtained from the interpretation of 104 mammograms by 10 radiologists. A cluster of microcalcifications was present in all cases; 46 cancers and 58 benign lesions were confirmed at biopsy. The effect of a computer aid was tested; it provided an estimate of the likelihood that microcalcifications were due to a malignancy. Sensitivity and specificity results were based on the radiologists' recommendations for biopsy or follow-up. The ROC curves were based on the radiologists' diagnostic confidence.

0.5, which represents no apparent accuracy (diagnoses corresponding to random chance alone), and 1.0, which represents perfect accuracy.

A histogram of interobserver agreement regarding clinical recommendations was constructed, and the κ statistic was computed. This histogram displayed the number of cases as a function of the number of observers in agreement. For the 10 observers, 11 patterns of agreement in the recommendations were possible; these patterns included 10 biopsy recommendations, nine biopsy and one follow-up recommendations, eight biopsy and two follow-up recommendations, and so on. For this analysis, we compared the recommendations for biopsy (option a or b in Table 1) versus those for follow-up (option c or d in Table 1), because this is the most important clinical decision. Separate histograms were constructed for cancers and benign lesions. Separate histograms were also constructed for attending radiologists, residents, and all radiologists. The histograms were similar for the three observer groups; we report only the summary histogram of all radiologists combined.

The κ statistic is widely used as a measure of agreement (16). It reflects the proportion of agreement after the proportion of agreement that can be attributed to chance alone is subtracted (17). κ equals 1 for perfect agreement, and κ equals 0 when the agreement can be attributed to chance alone. We computed the multireader κ value (18) and estimated the 95% CIs by using the bootstrap method.

Using the definitions by Elmore et al (1), we defined substantial disagreement as a situation in which one radiologist recommended biopsy (option a or b in Table 1) and another recommended routine follow-up (option d in Table 1) in the same case (short-term follow-up was excluded from this particular analysis to emphasize extremes in decision making). Pairwise and per-patient frequencies of substantial disagreement were calculated. The pairwise frequency was the occurrence of substantial disagreement in all recommendation pairs (ie, recommendations made by two different observers in the same case). The total number of recommendation pairs was equal to the following: [number of cases × number of readers × (number of readers − 1)]/2. For 10 observers, there were a total of (104 × 10 × 9)/2, or 4,680, recommendation pairs. The per-patient frequency was the fraction of total cases (ie, 104 cases) in which different observers simultaneously recommended at least one biopsy procedure and at least one routine screening procedure. Because of the large differences in the denominators, the pairwise frequency tended to produce a low estimate, and the per-patient frequency tended to produce a high estimate of the substantial disagreement; neither was clinically accurate, because it was unlikely that 10 radiologists would have independently evaluated the case in clinical practice. Because the true frequency of substantial disagreement was expected to be between the pairwise and per-patient frequencies and because, to our knowledge, no single accurate measure is known, we report both pairwise and per-patient results, as Elmore et al (1) did.

RESULTS

Effect of the Computer Aid on Sensitivity, Specificity, and ROC Curves

Sensitivity and specificity data and summary ROC curves are shown in Figure 1. The ranges and averages of the sensitivity, specificity, and positive predictive values are shown in Table 2. For the group of all observers (n = 10), without the computer aid there was a range of 35% in sensitivity and 44% in specificity. When the computer aid was used, the range in sensitivity was reduced to 26%, but the range in specificity remained 45%. Results for the groups of attending radiologists (n = 5) and residents (n = 5) were similar (Table 2). The average sensitivity, specificity, and positive predictive values increased significantly with the computer aid (8). Table 3 lists the A_z values. The SD of A_z values was reduced from 0.056 to 0.030, or 46%, with the computer aid.

Effect of Computer Aid on Agreement in Recommendations

The histogram of interobserver agreement (Fig 2) provides detailed information concerning the extent of agreement, for both cancers and benign lesions, and the changes as a result of the computer aid. With the computer aid, complete agreement among all 10 radiologists was achieved in 20 (43%) cancer cases. Agreement in benign cases had a broader distribution. Highlights of Figure 2 are summarized in Table 4. Without the computer aid, complete agreement by all observers on a correct recommendation (biopsy for cancers and follow-up for benign lesions) occurred in nine cases (nine malignant and no benign lesions). With computer aid, the

TABLE 2
Effect of CAD on Sensitivity, Specificity, and Positive Predictive Values

Value	All Observers (n = 10)		Attending Radiologists (n = 5)		Residents (n = 5)	
	Without Aid	With Aid	Without Aid	With Aid	Without Aid	With Aid
Sensitivity (%)						
Range	52–87	74–100	52–87	74–96	61–87	74–100
Average ± SD	74 ± 11	87 ± 9	75 ± 14	88 ± 9	72 ± 10	87 ± 10
Specificity (%)						
Range	9–53	22–67	9–53	28–60	19–52	22–67
Average ± SD	32 ± 15	42 ± 15	29 ± 18	42 ± 15	34 ± 14	42 ± 16
Positive predictive value (%)						
Range	42–52	51–64	42–51	51–60	43–52	51–64
Average ± SD	46 ± 3	55 ± 4	46 ± 4	55 ± 4	47 ± 3	55 ± 5

Note.—P values (Student t test) for the all observers, attending radiologists, and residents, respectively, were as follows: sensitivity, <.001 (t = 5.2, df = 9), .03 (t = 3.3, df = 4), and .02 (t = 3.7, df = 4); specificity, .003 (t = 4.1, df = 9), .007 (t = 5.0, df = 4), and .14 (t = 1.8, df = 4); and positive predictive value, <.001 (t = 11.9, df = 9), <.001 (t = 9.3, df = 4), and .002 (t = 7.7, df = 4).

TABLE 3
Effects of CAD on A_z

Observers	A_z without Aid	A_z with Aid
Attending radiologists		
A	0.64 (0.055)	0.76 (0.046)
B	0.60 (0.056)	0.76 (0.047)
C	0.72 (0.051)	0.77 (0.046)
D	0.54 (0.056)	0.72 (0.050)
E	0.59 (0.057)	0.79 (0.043)
Average	0.62 (0.067)	0.76 (0.028)
Radiology residents		
F	0.53 (0.056)	0.70 (0.050)
G	0.63 (0.055)	0.77 (0.045)
H	0.60 (0.057)	0.73 (0.049)
I	0.66 (0.054)	0.75 (0.047)
J	0.64 (0.056)	0.79 (0.044)
Average	0.61 (0.050)	0.75 (0.036)
All, average	0.61 (0.056)	0.75 (0.030)

Note.—Data in parentheses are SDs.

complete agreement on a correct recommendation increased to 26 cases (20 malignant and six benign lesions). Conflicting recommendations in which the minority consisted of more than 20% (ie, three to five of 10 observers or two of five observers) of the total observers occurred in 43 cases without aid; this number was reduced to 28 cases with the computer aid. Use of the computer aid improved agreement and reduced the occurrence of conflicting recommendations in all data categories (P < .05 with one exception of P = .07 in decreasing conflicting recommendations among residents; McNemar χ^2 test).

κ values are shown in Table 5. The results were consistent among the three observer groups (all radiologists, attending radiologists, and residents). Although the residents' κ values were smaller than those of the attending radiologists, the differences were not statistically significant (P > .05). In all three observer groups, use of the computer aid improved agreement from fair to moderate (on the ordinal scale where a κ value of 0.21–0.40 represents fair agreement beyond chance, and 0.41–0.60 represents moderate agreement beyond chance [19]). All improvements were statistically significant (P < .05).

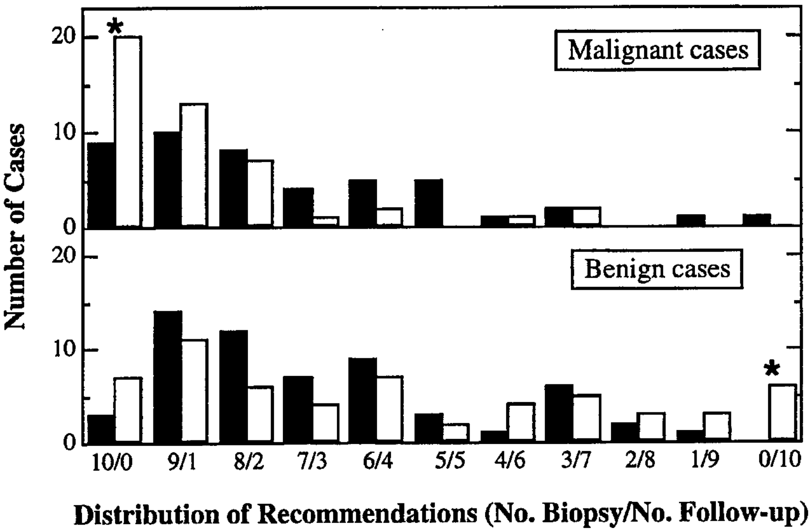

Figure 2. Histograms show the effect of CAD on the agreement in the recommendations for clinical management that were made by the 10 radiologists. Recommendations for biopsy versus any type of follow-up were made after the radiologists independently interpreted mammograms that depicted clustered microcalcifications. The computer aid provided an estimate of the likelihood that the microcalcifications were due to a malignancy. Black bars = without the computer aid, white bars = with the computer aid, and * = ideal situation of complete agreement in the correct recommendation.

Effect of Computer Aid on Substantial Disagreement in Recommendations

Interobserver agreement implicitly quantifies disagreement, but it does not distinguish between minor disagreements and completely incompatible diagnoses. Substantial disagreements represent contradictory diagnoses that can potentially cause greater confusion for the referring physicians and patients. Figure 3 shows the pairwise and per-patient frequencies of substantial disagreements. For recommendations made by attending radiologists without aid, the pairwise frequency of contradiction was 7%, and the per-patient frequency of contradiction was 23%. The frequencies were higher among residents: The pairwise frequency was 19%, and the per-patient frequency was 51%. Use of the computer aid reduced all occurrences of substantial disagreements. The reductions averaged 63% among attending radiologists and 28% among residents. The reduction was statistically significant for all cases combined and for cancers alone ($P < .04$ and $\chi^2 > 4.33$, with one exception of $P = .052$ and $\chi^2 = 3.77$ for residents and cancers alone; McNemar χ^2 test [The degree of freedom for the McNemar χ^2 test is always 1.]). The reduction was not statistically significant for benign cases alone ($P > .08$, $\chi^2 < 3.00$; McNemar χ^2 test).

TABLE 4
Effect of CAD on Agreement of Clinical Recommendations

Cases	All Observers (n = 10)		Attending Radiologists (n = 5)		Residents (n = 5)	
	Without Aid	With Aid	Without Aid	With Aid	Without Aid	With Aid
Complete agreements*	13 (13)	33 (32)	38 (37)	53 (51)	27 (26)	46 (44)
Correct recommendations	9	26	17	39	18	32
Missed cancers	1	0	1	1	3	0
Conflicting recommendations†	43 (41)	28 (27)	31 (30)	18 (17)	31 (30)	20 (19)

Note.—Data are the number of cases (n = 104). Data in parentheses are percentages.

* Complete agreement was defined as a situation in which all observers made the same recommendation for either biopsy or follow-up. P values (McNemar χ^2 test) for all observers, attending radiologists, and residents, respectively, were <.001 (χ^2 = 14.29), .05 (χ^2 = 3.95), and <.001 (χ^2 = 10.94).

† Conflicting recommendations were considered to occur when the minority opinion was held by more than 20% of the observers. P values (McNemar χ^2 test) for all observers, attending radiologists, and residents, respectively, were .02 (χ^2 = 5.49), .03 (χ^2 = 4.57), and .07 (χ^2 = 3.27).

TABLE 5
Effect of CAD on Agreement

Observers	κ without Aid	κ with Aid	Improvement*
All	0.19 (0.13, 0.28)	0.41 (0.32, 0.51)	0.22 (0.11, 0.33)
Attending radiologists	0.21 (0.13, 0.32)	0.44 (0.32, 0.56)	0.23 (0.07, 0.38)
Residents	0.17 (0.08, 0.28)	0.37 (0.26, 0.48)	0.19 (0.05, 0.33)

Note.—Data in parentheses are 95% CIs obtained by using the statistical method of bootstrapping with 10,000 repetitions.

* $P < .05$ in all improvement categories.

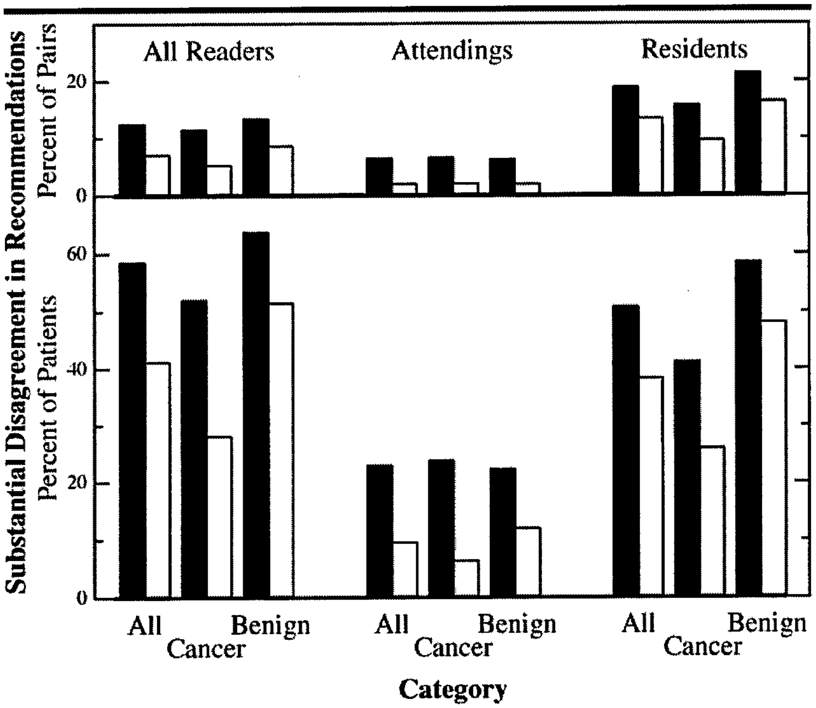

Figure 3. Histograms show the effect of CAD on substantial disagreements in clinical recommendations (ie, biopsy vs routine screening). Data shown are pairwise (top) and per-patient (bottom) frequencies. Pairwise frequencies were calculated from all pairs of recommendations made by two different radiologists. Per-patient frequencies were calculated from the total number of cases in which the recommendations were made by multiple radiologists (n = 5 for attending radiologists, n = 5 for residents, and n = 10 for all readers). Black bars = without the computer aid, and white bars = with the computer aid.

DISCUSSION

Interpretation of Results

To our knowledge, there is no single measure of agreement that can be universally used to quantify interpretation variability. Although κ is widely used as a quantitative measure of agreement, it is not without limitations when the findings of different studies are compared (20). More important, there is no explicit relationship between κ statistic and ROC analysis; the latter is often used to quantify diagnostic accuracy. Therefore, we extended our calculations beyond determining the κ value to three separate but related analyses. First, we calculated the variability that is evident in the ROC summary indices. This analysis could serve as a direct link between the calculation of variability and the calculation of diagnostic accuracy by means of ROC analysis. Second, we calculated the κ value and the pattern of agreement. Third, we assessed variability from the points of view of the referring physician and the patient by using a calculation in the literature (1). Each of the three analyses addressed a different aspect of variability, and together they helped to define its magnitude and the ability of CAD to help reduce the variability.

TABLE 6
Effect of CAD on Observer Variability

			Average Accuracy*		SD	
Study	Task of Interpretation	Accuracy Index	Without Aid	With Aid	Without Aid	With Aid
Getty et al (5)†	Differentiation of malignant and benign breast lesions on mammograms	A_z	0.832	0.880	0.046	0.020
Chan et al (21)‡	Detection of clustered microcalcifications on mammograms	A_z	0.924	0.953	0.044	0.041
Kegelmeyer et al (22)†	Detection of spiculated masses on mammograms	Sensitivity/specificity	0.806/0.954§	0.903/0.969§	0.051/0.054	0.048/0.039
Jiang et al (8)†	Differentiation of malignant and benign clustered microcalcifications on mammograms	A_z	0.614	0.755	0.056	0.030
Chan et al (23)†‖	Differentiation of malignant and benign masses on mammograms	A_z	0.860/0.912	0.900/0.947	0.044/0.026	0.037/0.020
Kobayashi et al (24)‡	Detection of lung nodules on chest radiographs	A_z	0.894	0.940	0.053	0.027
Difazio et al (25)‡	Detection of temporal change on chest radiographs	A_z	0.887	0.984	0.048	0.011
Monnier-Cholley et al (26)‡	Detection of interstitial infiltrates in chest radiographs	A_z	0.948	0.970	0.034	0.024
Ashizawa et al (27)‡	Differential diagnosis of interstitial lung disease on chest radiographs	A_z	0.826	0.911	0.038	0.024

Note.—Number in parentheses is the reference number.
* Unless otherwise specified, $P < .05$ for all improvements in the accuracy index, as reported in the original article.
† Data were derived from data in the reference.
‡ Data were taken directly from the reference.
§ $P > .05$ for all improvements in the accuracy index, as reported in the original article.
‖ Data are reported for reading condition 1/reading condition 2.

Our analysis revealed that there is considerable variability in the interpretation of mammograms by radiologists; this finding is consistent with that of other studies (1–4). We found similar or poorer agreement between radiologists, compared with the results of Elmore et al (1). Elmore et al reported a per-patient substantial-disagreement frequency of 25%, which is similar to our result of 23% for attending radiologists. However, our κ values were generally lower; these values suggested poorer agreement. This result may have been caused by differences in the calculation of κ values (ie, averaging two-reader κ values [1] vs calculating multireader κ values); also, our study (8) did not include cases that were not evaluated at biopsy. To increase the statistical power of the study by enhancing the proportion of cases that were difficult to diagnose, only abnormal cases that had biopsy confirmation were used in our study (11). These difficult cases can be presumed to generate more variability in interpretation. We found a range of 35% in sensitivity and a range of 44% in specificity. These results agree with the ranges of 53% in sensitivity and 45% in specificity reported by Beam et al (2), who studied the results of 108 radiologists.

Two sources can potentially generate variability in the interpretation of mammograms. First, variations in diagnostic accuracy (ie, variations in radiologists' abilities to correctly diagnose cancerous and cancer-free lesions) may be a primary source of variability. A_z values vary as a result of this variation. Second, a radiologist's selection of a decision threshold that defines a positive diagnosis in his or her interpretation can also produce variability (6). A decision threshold is necessary in all binary diagnostic tasks, and its selection is influenced by a radiologist's perception of disease prevalence and the benefits and costs associated with correctly diagnosing the disease (14). Although selection of different thresholds causes sensitivity and specificity to vary simultaneously and in opposite directions, such variations are not caused by and do not represent variations in diagnostic accuracy (6). Selection of the different thresholds does not cause A_z values to vary because an ROC curve, for which A_z is a summary index, depicts all of the tradeoffs available as the threshold is varied. Therefore, selection of the decision threshold is an issue that is separate from the variation in diagnostic accuracy as quantified with A_z values.

Our results (Fig 1) showed that the sensitivity and specificity data points were on or near the average ROC curves; these results indicated that much of the variation in sensitivity and specificity was caused by the use of different decision thresholds during interpretation and not by variations in diagnostic accuracy. This is consistent with the interpretation of D'Orsi and Swets (6) of the results of Elmore et al (1). As one might expect, the similarity in the ranges of sensitivity and specificity with and without use of a computer aid indicated that CAD had little influence on the radiologists' choices of decision thresholds, since CAD is not expected to influence the radiologists' perception of disease prevalence and the benefits and costs associated with correctly diagnosing the disease. The improvement in accuracy achieved with CAD is a result of the radiologists being able to improve their performance, as reflected with a different (higher) ROC curve. Moreover, compared with the without-aid data, the decreased dispersion of the with-aid sensitivity and specificity data points from the average ROC curve shows that CAD helped the radiologists to interpret mammograms with a more uniform, as well as higher, level of accuracy. Therefore, although CAD caused little change in the ranges of sensitivity and specificity (which in our study appear to have been determined largely by the radiologists' choice of decision thresholds), our results showed that CAD helped the radiologists to reduce variation in their diagnostic accuracy.

We analyzed data for attending radiologists and residents both in aggregate and in two separate groups, and we found that the results were similar except in the frequencies of substantial disagreement (Fig 3). We believe the residents' data are clinically relevant because the majority of recent residents and fellows who go into private practice are assigned to reading screening mammograms. Although on may interpret data from attending radiologists and residents differently, inclusion or exclusion of the residents' data did not alter the findings of this study.

Comparison with Other Observer Study Data

We compared our findings with those of eight other investigations (5,21–27) of the effects of CAD on observer performance. By re-analyzing the results of these studies, we deduced general conclusions that are not limited to a particular computer aid or imaging task, as these were different in each of the studies; rather, our conclusions pertain to CAD in general. We used the accuracy indices (A_z in all studies except one) and corresponding SDs that were reported in the original investigations as measures of diagnostic accuracy and variability. The results (Table 6) showed that accuracy was always higher and that its SD was always smaller when a computer aid was used; these results indicated that accuracy was consistently improved and that variability was consistently reduced when a computer aid was used. Although these studies were not specifically designed to measure the effect of a computer aid on reader variability, the clear trend of a reduction in reader variability in all of these nonuniformly designed studies indicates that the reduction is likely a consequence of, rather than a coincidental finding with, use of a computer aid.

The ability of CAD to improve diagnostic accuracy is conceptually similar to double reading by two radiologists (28), in which gains in accuracy are expected if two radiologists are able to complement each other (29). Several investigators (5,8,21–27) suggest that the clinical role of CAD might be to serve as a less expensive alternative to double reading by radiologists. However, to our knowledge, the ability of CAD to reduce variability has not been previously investigated, and we present the first evidence. We believe that the computer aid provides a reference point, much as reading with a skilled partner does. In clinical practice, the variability of the second human reader is one of the major problems that prohibits widespread use of this technique, despite its promised advantages. Because the computer aid is used independent of the radiologists' interpretations of the mammograms, it can serve as a reference reader that is completely immune to human variability. This could be a unique advantage of CAD when it is compared with other approaches for reducing variability that depend on radiologists' interpretations, which are subject to the inherent variation in human perception and decision making. CAD also eliminates or reduces the need for arbitration or reconciliation between differing opinions when two human readers disagree, because the course of action is ultimately determined by the radiologist using the added opinion of the computer output. The final clinical decision remains in the hands of a single human reader, and studies (28–31) have consistently shown that computer-assisted readers perform at a higher level, with improvement comparable to or exceeding that seen in traditional double-reader studies.

Impediments to the clinical use of CAD include the radiologic community's underestimation of the extent of individual variability in daily practice and the effects that missing important low-prevalence events or overreacting to common benign conditions has on screening. Recent studies have focused attention on these problems and have created an appreciation of the need for more standardization of the observer's role in the screening process.

In summary, a CAD joint reading could promote agreement and eliminate some of the extreme or erroneous diagnostic opinions. Both of these outcomes are highly desirable in the medical, social, and economic contexts of breast cancer screening in an asymptomatic population.

Two major conclusions can be drawn from our data and our findings from analysis of nine independent observer-performance studies (5,8,21–27): CAD can improve diagnostic performance, and CAD can simultaneously reduce interpretation variability. These beneficial effects are possible because CAD can help radiologists to avoid performing biopsy in benign lesions, while it increases, rather than decreases, the number of correct diagnoses of cancers. The second capability is a substantial enhancement to the known potential of CAD, which has been demonstrated in several studies. Our findings suggest that if CAD is incorporated into clinical radiology, improvements in both accuracy and consistency in image interpretation can be expected. Patients and referring physicians would agree that both of these goals are highly desirable. These goals support the intention of the Breast Imaging Reporting and Data System lexicon introduced by the American College of Radiology and the Mammography Quality Standards Act, that is, to improve the daily practice and results of breast cancer screening by fostering more uniform interpretations.

Acknowledgments: We thank Charles E. Metz, PhD, for reviewing the manuscript and for use of his LABROC4 program that generated the fitted ROC curves. We thank Carl J. Vyborny, MD, PhD, for his insightful comments.

References

1. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994; 331:1493–1499.
2. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. Arch Intern Med 1996; 156:209–213.
3. Schmidt RA, Newstead GM, Linver MN, et al. Mammographic screening sensitivity of general radiologists. In: Karssemeijer N, Thijssen M, Hendriks J, van Erning L, eds. Digital mammography. Dordrecht, the Netherlands: Kluwer Academic, 1998; 383–388.
4. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst 1998; 90:1801–1809.
5. Getty DJ, Pickett RM, D'Orsi CJ, Swets JA. Enhanced interpretation of diagnostic images. Invest Radiol 1988; 23:240–252.
6. D'Orsi CJ, Swets JA. Variability in the interpretation of mammograms (letter). N Engl J Med 1995; 332:1172.
7. Doi K, MacMahon H, Katsuragawa S, Nishikawa RM, Jiang Y. Computer-aided diagnosis in radiology: potential and pitfalls. Eur J Radiol 1999; 31:97–109.
8. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 1999; 6:22–33.
9. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol 1998; 5:155–168.
10. Sickles EA. Breast calcifications: mammographic evaluation. Radiology 1986; 160: 289–293.
11. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
12. Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York, NY: Academic Press, 1982.
13. Jiang Y, Nishikawa RM, Wolverton DE, et al. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology 1996; 198:671–678.
14. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
15. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988; 240:1285–1293.
16. Fleiss JL. Statistical methods for rates and

proportions. 2nd ed. New York, NY: Wiley, 1981.

17. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20:37–46.
18. Fleiss JL. Measuring nominal scale agreement among many raters. Psych Bull 1971; 76:378–382.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159–174.
20. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987; 126:161–169.
21. Chan HP, Doi K, Vyborny CJ, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis. Invest Radiol 1990; 25:1102–1110.
22. Kegelmeyer WP Jr, Pruneda JM, Bourland PD, Hillis A, Riggs MW, Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994; 191: 331–337.
23. Chan HP, Sahiner B, Helvie MA, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999; 212:817–827.
24. Kobayashi T, Xu XW, MacMahon H, Metz CE, Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology 1996; 199:843–848.
25. Difazio MC, MacMahon H, Xu XW, et al. Digital chest radiography: effect of temporal subtraction images on detection accuracy. Radiology 1997; 202:447–452.
26. Monnier-Cholley L, MacMahon H, Katsuragawa S, Morishita J, Ishida T, Doi K. Computer-aided diagnosis for detection of interstitial opacities on chest radiographs. AJR Am J Roentgenol 1998; 171: 1651–1656.
27. Ashizawa K, MacMahon H, Ishida T, et al. Effect of an artificial neural network on radiologists' performance in the differential diagnosis of interstitial lung disease using chest radiographs. AJR Am J Roentgenol 1999; 172:1311–1315.
28. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191: 241–244.
29. Metz CE, Shen JH. Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis. Med Decis Making 1992; 12:60–75.
30. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol 1996; 3:891–897.
31. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Doi K. Comparison of independent double reading and computer-aided diagnosis (CAD) for the diagnosis of breast lesions (abstr). Radiology 1999; 213(P): 323.

Components-of-Variance Models for Random-Effects ROC Analysis:

The Case of Unequal Variance Structures Across Modalities[1]

Sergey V. Beiden, PhD, Robert F. Wagner, PhD, Gregory Campbell, PhD, Charles E. Metz, PhD, Yulei Jiang, PhD

Rationale and Objectives. Several of the authors have previously published an analysis of multiple sources of uncertainty in the receiver operating characteristic (ROC) assessment and comparison of diagnostic modalities. The analysis assumed that the components of variance were the same for the modalities under comparison. The purpose of the present work is to obtain a generalization that does not require that assumption.

Materials and Methods. The generalization is achieved by splitting three of the six components of variance in the previous model into modality-dependent contributions. Two distinct formulations of this approach can be obtained from alternative choices of the three components to be split; however, a one-to-one relationship exists between the magnitudes of the components estimated from these two formulations.

Results. The method is applied to a study of multiple readers, with and without the aid of a computer-assist modality, performing the task of discriminating between benign and malignant clusters of microcalcifications. Analysis according to the first method of splitting shows large decreases in the reader and reader-by-case components of variance when the computer assist is used by the readers. Analysis in terms of the alternative splitting shows large decreases in the corresponding modality-interaction components.

Conclusion. A solution to the problem of multivariate ROC analysis without the assumption of equal variance structure across modalities has been provided. Alternative formulations lead to consistent results related by a one-to-one mapping. A surprising result is that estimates of confidence intervals and numbers of cases and readers required for a specified confidence interval remain the same in the more general model as in the restricted model.

Key Words. Receiver operating characteristic (ROC); components-of-variance; jackknife; bootstrap.

The field of random-effects receiver operating characteristic (ROC) analysis has made important advances during the past decade. Its major applications include the assessment of modalities for diagnostic imaging and computer-assisted diagnosis (CAD) and the comparison of competing diagnostic modalities. A particularly important paradigm is the multiple-reader, multiple-case (MRMC) approach in which every reader reads every patient case. This is the so-called reader study that allows for a proper accounting of both reader and case variance and thus provides estimates of uncertainties of ROC parameters that are said to be "generalizable to a population of readers as well as to a population of cases." This paradigm was first modeled by Swets and Pickett in 1982 (1). Dorfman, Ber-

Acad Radiol 2001; 8:605–615

[1] From the Offices of Science and Technology (HFZ-142) (S.V.B., R.F.W.) and Surveillance and Biometrics (HFZ-542) (G.C.), Center for Devices and Radiological Health, Food and Drug Administration, Rockville, MD 20857; and the Department of Radiology, Rossmann Laboratories, University of Chicago, Ill (C.E.M., Y.J.). Received July 24, 2000; revision requested December 14; revision received March 4, 2001; accepted March 5. S.V.B. supported in part by an appointment to the Research Participation Program at the Center for Devices and Radiological Health administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. C.E.M. supported by grant R01-GM57622 from the National Institutes of Health and by a University of Chicago contract with the University of Iowa (Donald D. Dorfman, PI) under grant R01-CA62362 from the National Institutes of Health. Y.J. supported by U.S. Army Medical Research and Materiel Command grant DAMD17-00-1-0197. **Address correspondence to** R.F.W.

baum, and Metz (DBM) later provided a more flexible theoretical and also more practical solution to the MRMC problem (2). Their use of a general linear model together with "jackknife" resampling allowed the application of standard analysis-of-variance (ANOVA) techniques. Their approach and several alternatives were discussed at a 1993 symposium, and the proceedings were published in a supplement to this journal (3).

Beiden, Wagner, and Campbell (BWC) have recently provided a review of some of the issues in random-effects ROC analysis, together with an alternative solution to the MRMC problem (4). The BWC analysis includes not only the estimation of uncertainties in performance estimates in the MRMC paradigm but also a method to uniquely decompose these uncertainties into contributions in a components-of-variance model (2,4,5). These components are referred to as the "variance structure" of the problem and include the case variability, the reader variability, various interactions among cases, readers, and modalities and, finally, experimental replication error. The BWC alternative to previous solutions involves the analysis of a set of population experiments in terms of the model components. In any realistic clinical context, such population experiments are not possible. The practical solution is to replace the set of population experiments with the set of corresponding bootstrap resampling experiments on the available finite data set. This leads to a system of linear equations that may be solved for estimates of the components of variance (ie, the sources of randomness). In turn, one then obtains estimates of the confidence intervals of interest, as well as the ability to size a pivotal study from a pilot study.

In the previous work, we followed the model and assumption of DBM, namely, that the reader and case variances and their interaction for one modality are so similar to those for the other modality that they can be assumed to be equal. A central goal of CAD and other evolutions in imaging technology, however, is to create new modalities that will outperform older ones—in ways that include reducing the magnitude of these components of variance. A comparison of the performance of such new and older technologies will therefore require a more general model.

In the present article, we extend our previous work to the more general case of unequal variance structures across two modalities under comparison: We will show how to solve for estimates of the variance structure for this more general MRMC paradigm. In the next section, we present one formulation of a solution to this estimation problem. An alternative formulation is presented in the Appendix. Our analysis is applied to the study of Jiang et al (6), in which unaided readers of suspicious mammographic clusters of calcifications were compared with readers who used a CAD modality as an adjunct. In a companion article (7), we analyze the uncertainties in the estimates of the variance structure.

MATERIALS AND METHODS

Following DBM (2), we analyze the MRMC paradigm within the framework of a general multivariate linear model for ROC parameter estimates. We will use the ROC area parameter, A_z (dropping the z for simplicity), to exemplify the model; the model is nevertheless applicable to any other ROC model parameter or accuracy index. For completeness, we repeat the multivariate linear model for an ROC accuracy index, A, used by DBM:

$$\begin{aligned} A_{ijkn} = \mu_i + r_j + c_k + (mr)_{ij} \\ + (mc)_{ik} + (rc)_{jk} + (mrc)_{ijk} + z_{ijkn}, \end{aligned} \qquad (1)$$

where i indicates a particular imaging modality, j denotes a particular image reader, k is a particular case sample, and n is a particular replication of the experiment. (The index for case sample, k, is included in this model because DBM studied jackknife pseudo-values.) The term μ_i represents the contribution of modality i to the expected value of the accuracy index, while the remaining terms are independent zero-mean random variables. The terms with a single index are the reader and case contributions to the variability, with variances σ_r^2 and σ_c^2, respectively. The terms with two subscripts represent the two-way interactions between modality and reader, modality and case, and reader and case, with variances σ_{mr}^2, σ_{mc}^2, and σ_{rc}^2, respectively. The term with three subscripts represents the three-way interaction among modality, reader, and case, with variance σ_{mrc}^2. The last term is a pure error term in experimental reproducibility, with variance σ_z^2. For the case where multiple-reader experiments are conducted but readers do not independently repeat their readings, the last two terms, with variances σ_{mrc}^2 and σ_z^2, are inseparable, and we combine them into a single term with variance σ_ε^2.

A major distinction in applications of this model is that between random and fixed factors. A random factor is one that—on replication of the experiment—is drawn independently from a specified population; a fixed factor is one that remains unchanged on replication. As written in Equation (1), modalities are considered fixed factors, while readers and cases are random factors. Finally, we note that it will not be necessary for the present article to invoke assumptions of normality.

The model of Equation (1) assumes that the variance structure is homogeneous across modalities. Our present interest, however, is the case in which this structure changes across modalities. A parsimonious model for this case and the application where readers do not independently repeat their readings can be obtained by making the reader, case, and reader-by-case interaction terms a function of modality (i), respectively, $r_j(i)$, $c_k(i)$, $(rc)_{jk}(i)$. It can be written as

$$A_{ijkn} = \mu_i + r_j(i) + c_k(i) + (mr)_{ij} + (mc)_{ik} + (rc)_{jk}(i) + \epsilon_{ijk}. \qquad (2)$$

The two-way interaction terms involving modality m and readers r (or cases c) carry information related to the reader (or case) correlation across modalities; they do not require generalization. (All else being equal, the interaction strength is higher when the correlation is lower, and vice versa.) However, for the case where readers independently repeat their readings, the three-way interaction term also would not be made a function of modality, but the final term in Equation (1) would be. An alternative formulation, described in the Appendix, generalizes Equation (1) in a different way.

The variances produced by any linear model, such as Equation (2), and that contribute to observations over repeated experiments depend on which factors are held fixed and which are sampled randomly from a population when a particular ROC experiment is repeated. In reference 4 we showed that, for the equal-variance model considered there, it is possible to perform six population experiments, chosen from the family of 32 considered by Roe and Metz (5), that would allow one to solve for the six variance components in Equation (1), combining the final two components as just described. In the present work, we extend this approach to solve for nine components in the new model, using nine equations.

We use the notation of Roe and Metz (5), where variables to the left of the vertical bar in the subscript of an accuracy index are random factors, while those to the right are fixed factors. For example, suppose we consider replications of the experiment where readers R as well as cases C are drawn randomly from the population but the modality M is a fixed factor. All six variance components for a given modality contribute to the observed variance in this experiment. This is stated by the following expression, which, for the case of two modalities, provides two equations:

$$\mathrm{var}(A_{RC|M}) = \sigma_r^2(M) + \sigma_c^2(M) + \sigma_{mc}^2 + \sigma_{mr}^2 + \sigma_{rc}^2(M) + \sigma_\epsilon^2. \qquad (3)$$

When readers are also a fixed effect, the pure reader term and the modality-by-reader term do not contribute. That experiment and observed variance are given by

$$\mathrm{var}(A_{C|MR}) = \sigma_c^2(M) + \sigma_{mc}^2 + \sigma_{rc}^2(M) + \sigma_\epsilon^2, \qquad (4)$$

which also provides two equations when two modalities are being studied.

An experiment that is generally of most interest is the one in which two fixed modalities, M and M', are compared in terms of the ROC performance estimates obtained from randomly drawn reader and case samples. The population variance that is observed in that experiment can be calculated after subtracting two equations of the form of Equation (2) above:

$$\begin{aligned} A_{jk|1} - A_{jk|2} = {} & [\mu_1 - \mu_2] + [r_j(1) - r_j(2)] \\ & + [c_k(1) - c_k(2)] + [rc_{jk}(1) - rc_{jk}(2)] \\ & + [mr_{1j} - mr_{2j}] + [mc_{1k} - mc_{2k}] \\ & + [\epsilon_{jk|1} - \epsilon_{jk|2}]. \end{aligned} \qquad (5)$$

The first term in square brackets on the right-hand side is not a random variable, and so it contributes no variance. The variance of the next term in square brackets involves the correlation of $r_j(1)$ and $r_j(2)$. In the present model, we take these components to be different in magnitude but perfectly correlated, that is, $r_j(1) = \gamma_r r_j(2)$, where γ_r is a constant. (We treat the pure case and reader-by-case components similarly.) Thus,

$$\mathrm{var}[r_j(1) - r_j(2)] = [\sigma_r(1) - \sigma_r(2)]^2. \qquad (6)$$

This approach is consistent with the interpretation that the reader component was originally not a function of modality for the equal-variance model of Equation (1) and thus could be thought of as perfectly correlated across modalities in this special case to which the present model degenerates. More generally, of course, the reader variation may not be perfectly correlated across modalities. However, the flexibility to include arbitrary correlations of readers (or cases) across modalities is achieved in the general linear model as used here through the presence of the interaction terms. (For split-plot designs, however, where the readers [or cases] are drawn independently for the two modalities [8], the present model would be modified to set the reader [or case] correlation across modality to zero.)

By similar steps, the variance of the complete difference expressed by Equation (5) can then be written as

$$\begin{aligned}\mathrm{var}(A_{RC|M} - A_{RC|M'}) = {} & 2(\sigma_{mr}^2 + \sigma_{mc}^2 + \sigma_\epsilon^2) \\ & + (\sigma_r(M) - \sigma_r(M'))^2 \\ & + (\sigma_c(M) - \sigma_c(M'))^2 \\ & + (\sigma_{rc}(M) - \sigma_{rc}(M'))^2. \qquad (7)\end{aligned}$$

One similarly obtains the following results for the other experiments that are required in order to solve for all of the variance components in this model:

$$\mathrm{var}(A_{C|RM} - A_{C|R'M}) = 2(\sigma_{rc}^2(M) + \sigma_\epsilon^2), \qquad (8)$$

$$\begin{aligned}\mathrm{var}(A_{C|RM} - A_{C|RM'}) = {} & 2(\sigma_{mc}^2 + \sigma_\epsilon^2) \\ & + (\sigma_c(M) - \sigma_c(M'))^2 \\ & + (\sigma_{rc}(M) - \sigma_{rc}(M'))^2, \qquad (9)\end{aligned}$$

$$\begin{aligned}\mathrm{var}(A_{C|RM} - A_{C|R'M'}) = {} & \sigma_{rc}^2(M) + \sigma_{rc}^2(M') \\ & + 2(\sigma_{mc}^2 + \sigma_\epsilon^2) \\ & + (\sigma_c(M) - \sigma_c(M'))^2. \qquad (10)\end{aligned}$$

Notice that Equations (3), (4), and (8) each describe two independent experiments (M = 1 or 2). The system of nine equations represented by Equations (3), (4), and (7)–(10) then expresses nine observable variances as a multivariate quadratic equation in the square roots of nine variance components. These equations reduce to the linear expressions in our previous work (4) for the case where the variances are equal across modalities.

The left-hand sides of Equations (3), (4), and (7)–(10) are observables that are independent of any model. Thus, we may equate the right-hand sides just derived for the present model with the corresponding right-hand sides that follow from the model for the equal-variance case given in BWC (4). When necessary to distinguish between models, we shall use the presubscript A to refer to components in the BWC model of reference 4 and the presubscript B to refer to components in the model described up to this point in the present article. (Components of variance in the alternative formulation described in the Appendix will be denoted by a presubscript C. Otherwise in this article, the components will refer to the present model, model B.) For example, equating Equation (8) as written above to the corresponding version of this for the equal-variance case yields

$$[{}_B\sigma_{rc}^2(1) + {}_B\sigma_{rc}^2(2)]/2 + {}_B\sigma_\epsilon^2 = {}_A\sigma_{rc}^2 + {}_A\sigma_\epsilon^2. \qquad (11)$$

The average on the left-hand side of this equation results from the fact that in BWC (4) the observable quantity was taken to be the average over the fixed effect M, and thus we average over the two equations implied by Equation (8).

By repeating this exercise with Equations (4) and (3), and taking differences with Equation (11), two additional expressions parallel to Equation (11) can be found: one in which all versions of σ_c^2 replace the corresponding versions of σ_{rc}^2 and all versions of σ_{mc}^2 replace the corresponding versions of σ_ϵ^2, and another in which all versions of σ_r^2 replace the corresponding versions of σ_{rc}^2 and all versions of σ_{mr}^2 replace the corresponding versions of σ_ϵ^2. The complete parallel of these expressions with Equation (11) becomes apparent on recalling that σ_ϵ^2 includes σ_{mrc}^2.

Another set of relationships can be found by first performing a similar exercise on Equations (10) and (4). The difference of the two results yields

$${}_A\sigma_c^2 = {}_B\sigma_c(1){}_B\sigma_c(2). \qquad (12)$$

Similar results follow for the components σ_r^2 and σ_{rc}^2. These expressions will be useful as a check on the results below.

We now proceed as in our previous analysis (4) where, in practice, we replace a given population experiment with the corresponding bootstrap experiment. (Details of the statistical bootstrap are reviewed in reference 4, based on Efron [9] and Efron and Tibshirani [10].)

The nonlinear system, Equations (3), (4), and (7)–(10), can be solved for the unknown variance components by

Components of Variance in Equal- and Unequal-Variance Models (All Values × 10^{-4})

Variance Component	Equal Variance	Unequal Variance	
		Modality 1	Modality 2
c	7.31	7.66	6.98
r	7.78	17.84	3.39
rc	2.11	10.92	0.41
mc	4.43	4.42	. . .*
mr	7.72	4.48	. . .
mrc/ϵ	14.00	10.45	. . .

*Ellipses indicate no new parameter; these components do not split in the new model.

numerical iteration. We first write the variance components as a vector $\boldsymbol{\sigma}$, whose transpose, *T*, is

$$(\sigma)^T = (\sigma_c(1), \sigma_c(2), \sigma_r(1), \sigma_r(2), \sigma_{mc}, \sigma_{mr}, \sigma_{rc}(1), \sigma_{rc}(2), \sigma_\epsilon). \qquad (13)$$

Each of the nine equations is then rearranged such that the vector $\boldsymbol{\sigma}$ is on the left-hand side of the system; the right-hand side is then the remaining nonlinear operation on $\boldsymbol{\sigma}$, which we call *f*. The system can then be written as

$$\boldsymbol{\sigma} = f(\boldsymbol{\sigma}). \qquad (14)$$

We use a method of simple iteration to solve this system, with the initial estimate being taken to be the solution of the linear system that results when the two structures are equal. This system of quadratic equations can be shown to have only one physically meaningful solution set, and thus the problem is well defined.

RESULTS

Application to CAD

We use the study of CAD by Jiang et al (6) to exemplify this approach. These authors compared the performance of 10 radiologists—unaided versus with the aid of a computer-assist modality—reading mammograms from 104 patients with clustered microcalcifications. The truth state for these patients was established with biopsy (46 malignant, 58 benign cases). ROC analysis for individual readers and also their average performance within the MRMC paradigm and model of DBM were published in reference 6. Here we use the methods described above to solve for the components of variance in these MRMC experiments, both in terms of the previous model that assumes equal variance structure across modalities and in terms of the new model that does not make that assumption.

Components of Variance

In the Table, we present the components of variance according to our present analysis within the two models: the first model assuming equal variances across modalities and the second model assuming unequal variances. Here, the first modality (modality 1) refers to the combination of mammographic images and unaided readers; the second modality (modality 2) refers to the combination of mammographic images and readers aided by the computerized feature extraction, fusion, and rating of probability of cancer described in reference 6.

The following observations can be made from the Table. In the equal-variance model, the reader component of variance and the case component of variance have similar strengths. The patient component of variance, which can be interpreted as the range of case difficulty as represented by the finite sample, hardly changed when we went to the unequal-variance treatment. The reader component of variance, which can be interpreted as a range of reader ability, "splits" into two quite unequal components in the unequal-variance model. Without the assist of CAD, the reader component is now seen to be much greater than the case component; the addition of CAD is seen to reduce this component more than fivefold. The reader-by-case component also splits into two quite unequal components, with a more than 25-fold reduction after the addition of CAD. Larger values of this component imply that the range of sampled case difficulty depends on the particular reader (or by the symmetry of that component, that the range of reader skills depends on the case); smaller values imply less such dependence. Thus, we take this splitting to indicate that the addition of CAD in the study of reference 6 almost eliminated the dependence of the range of case difficulty on the particular reader in that study (or, symmetrically, that it almost eliminated the dependence of the range of reader skills on the case).

For later reference (companion article, reference 7) these results are shown graphically in Figure 1, together with error bars on the model results that represent ± 1 standard deviation. In the companion article (7), we provide the analysis of uncertainty in these results. In a few words, the error bars are obtained by using a resampling technique known as the jackknife-

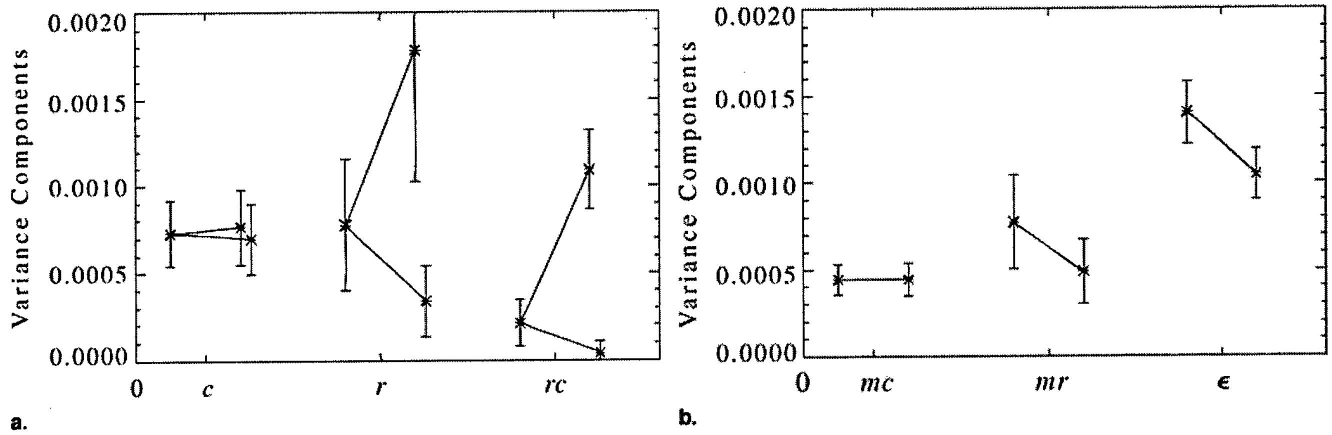

Figure 1. **(a)** Variance components *c* (case), *r* (reader), and *rc* (reader-by-case) in the present analysis of the study in reference 6. Vertical bars represent mean estimates, ± 1 standard deviation estimated with the method of reference 7. Unsplit components (to the left in each set of three) are estimated with the model of reference 4 (denoted model A). Splitting components (the pair to the right in each set) are estimated with model B of the present analysis (ie, not assuming equal variance structure across modalities). **(b)** Variance components *mc* (modality-by-case), *mr* (modality-by-reader), and ϵ (residual error) in the present analysis of the study in reference 6. These three components shift rather than split in going from model A of previous work to model B of the present article.

after-bootstrap (10), followed by linear propagation of variance for the model of equal variance structure or its modification for the model of unequal variance structure.

We note also that the model components in the unsplit model are indeed the geometric mean of the model components in the split model, consistent with the theoretical analysis of the model. Thus far, these observations are for the splitting components of the model. We now turn to the components that are not split in the new model.

In the new model, the reader-by-modality interaction changed to accommodate the new values of reader variance. There was little change in the case-by-modality interaction, as expected from the small change in the case component. Finally, the last component, or effective error term, is reduced when going to the new model. (The effective error term includes the contribution to the variance due to reader inconsistency that was called "jitter" in reference 11 and subsequent parlance.)

The shifts in the unsplit model terms that accompany the move to the more elaborate model can be accounted for by simple algebraic relationships. The change in the effective error term just noted, that is, the difference between the solution to the linear system, ${}_A\sigma^2_\epsilon$, and the corresponding solution to the quadratic system, ${}_B\sigma^2_\epsilon$, is simply related to the change in going from the solution to the linear system, ${}_A\sigma^2_{rc}$, to the solutions to the quadratic system, ${}_B\sigma^2_{rc}(1)$, ${}_B\sigma^2_{rc}(2)$. The relationship is found from Equations (11) and (12):

$$\begin{aligned}[{}_B\sigma^2_{rc}(1) + {}_B\sigma^2_{rc}(2)]/2 - {}_B\sigma_{rc}(1)\,{}_B\sigma_{rc}(2) &= [{}_B\sigma_{rc}(1) - {}_B\sigma_{rc}(2)]^2/2 \\ &= {}_A\sigma^2_\epsilon - {}_B\sigma^2_\epsilon. \qquad (15)\end{aligned}$$

That is, the shift in the nonsplitting component is the difference of the geometric mean and the arithmetic mean of the splitting components. Two additional expressions exactly parallel to Equation (15) can be found: one in which σ_c replaces σ_{rc} everywhere on the left-hand side and σ^2_{mc} replaces σ^2_ϵ on the right-hand side of Equation (15), and another in which σ_r replaces σ_{rc} everywhere on the left-hand side and σ^2_{mr} replaces σ^2_ϵ on the right-hand side of Equation (15). (Recall again that σ^2_ϵ contains σ^2_{mrc}.)

DISCUSSION

Inference and Experimental Design

An important consequence of the model and analysis above is that the present approach does not change the confidence intervals on the difference of ROC parameters between competing modalities, compared with our previous work (4). These confidence intervals are found from the single-bootstrap experiment represented by the left-hand side of Equation (7). The right-hand side of this equation is new in the present model, and thus the interpretation is new. The input to the equation represented by the left-hand side has not changed, however.

The new model also does not change the design of a pivotal study from results of a pilot study that was described in our previous analysis (4). In that analysis, the variance components *mc*, *mr*, and ϵ were the only contributors to the estimation of the numbers of cases and readers required for a specified confidence interval on the difference of ROC parameters between competing modalities, but in the present analysis there are nine contributions to that estimation task (right-hand side of Equation [7]). Although the former three terms may be reduced in the new model, inspection of Equation (15) and its analogues shows that this reduction is exactly offset by the remaining terms of Equation (7). Thus, the design of experiments according to the previous model and model parameters obtained in reference 4 is unchanged, if only the difference in performance between two modalities is of interest. However, the partitioning of the variances obtained in the present work provides additional insight for the entire family of possible experiments embraced in Equations (3), (4), and (7)–(10).

Since confidence intervals on differences between modalities and associated inferences based on them in our analysis do not change when going to the new model, it would be reasonable to expect that inferences based on an elaboration of the DBM analysis to the case of unequal variance structures across modalities would also remain unchanged. We have argued in reference 4 that our analysis is a distribution-free generalization of the approach of DBM. Since inferences based on this generalization remain unchanged when the variance structure is allowed to change across modalities, inferences based on an elaboration of DBM (ie, use of the jackknife rather than the bootstrap) might also be expected to remain unchanged. In the Appendix, we present an alternative to model B that contains no expressions nonlinear in the variance components and could thus be readily incorporated into the method of DBM. We refer to this alternative as model C. In the approach of the present article, inferences and design of experiments based on model C are identical to those based on model B, and they are thus identical to those based on BWC (4) when estimation of differences between modalities is the object of the experiment.

Generality of the Present Work

An anonymous reviewer (December 2000) has suggested that one could address the present problem by a natural extension of the DBM approach, using jackknifed pseudo-values with the PROC MIXED routines in the SAS software package (12). This will lead not only to estimates of the confidence intervals of interest but also to maximum-likelihood (ML) or residual (often called restricted) maximum-likelihood (REML) estimates of the variance components. The approach of using REML to obtain estimates of the variance components had also been mentioned to one of us (R.F.W.) previously (D.D. Dorfman, oral communication, 1999). We agree that this is indeed a reasonable alternative to the present approach, but it does not address the level of generality we seek here. We summarize this issue as follows.

The BWC approach (4), and its extension to the case of unequal variance structures as provided above, is built on the same general components-of-variance model used by DBM. However, it replaces the jackknife and ANOVA with a family of bootstrap resampling experiments and a corresponding system of equations that lead to explicit solutions for the variance components and confidence intervals of interest. It is thus a distribution-free approach, whereas classic ANOVA is based on the assumption of normality for all the components. (REML also requires assumptions for the relevant distributions.)

An additional feature of the present approach was cited in reference 4. The bootstrap includes not only the leave-one-out jackknife, but also more general leave-*X*-out terms where *X* is greater than one, among the other kinds of terms that sampling with replacement generates. For the case where the statistic of interest is linear, all of the terms that can contribute to the calculation of that statistic on a single-bootstrap pass are already included in the jackknifed data sets; this is not true of nonlinear statistics, that is, statistics that involve interactions between the data points two or more at a time (10). The nonparametric estimate of ROC area, for example, includes sums of rankings of data points two at a time (13,14) and thus falls into the latter category. Thus, the leave-one-out jackknife does not in general capture all of the information in the data regarding this statistic. Nevertheless, only small differences were found in reference 4 between the DBM and BWC methods for the variance structures and samples sizes studied there. Also, in our (unpublished) Monte Carlo simulations of bootstrap and jackknife estimates of variance for the nonparametric measure of ROC area, small differences between mean estimates were seen, but only when the number of patients per class was smaller than 25. This issue bears further investigation, including the case of parametric accuracy measures.

Finally, we emphasize a general point about the philosophy of the bootstrap made by Efron and Tibshirani (10). The empirical distribution function is the nonparametric ML estimate of the population distribution. In this sense, the nonparametric bootstrap provides "nonparametric ML estimates" in the language of reference 10, or "distribution-free" ML estimates in language that we and others prefer. The system of equations used here to propagate those estimates back into estimates of the variance components will thus also lead to distribution-free ML estimates. (This follows since the ML estimate of a function of a parameter of interest is that function of the ML estimate of the parameter.) For all of the above points, we would argue that the approach of reference 4 and its present extensions are the most general proposed so far for the family of problems under consideration here.

CONCLUSIONS

The present approach to random-effects ROC analysis extends our previous work (4) to the case where the variance structure may change across modalities. An example comparing unaided readers with readers assisted by CAD showed that both the reader and the reader-by-case components of variance were greatly reduced after the addition of CAD. These results are consistent with previous expectations regarding that study (15), but such results had not been previously isolated quantitatively.

Several comments regarding the future are in order. The present model provides a quantitative framework for interpreting the variability in MRMC studies in terms of a model of the components of that variability. It may thus offer the opportunity to contribute to the solution of several outstanding problems in the field of medical image science. The first of these is the connection between physical performance measurements on diagnostic imaging systems, that is, measurements of "image quality," and measures of clinical outcome such as the ROC curve (16,17). The variability observed at present in mammographic imaging (18), to take just one example, may mask the gains to be expected from evolution of the physical performance of mammographic imaging systems. The present approach may make it possible to peel back this mask with an efficient clinical experimental design.

The ability to isolate the contributions to variability in performance that arise from the reader from those that arise from the patient and the imaging system opens up new possibilities for imaging system optimization. The professional community of radiologists may be better able to quantitatively measure and fine-tune their training of readers, while the professional community of physicists and engineers of imaging systems may be better able to fine-tune their system designs, each with the appropriate focus and emphasis.

Finally, the emergence of the field of computer-assisted reading of images adds another layer of complexity to the problem of assessing diagnostic imaging modalities. The present work may contribute toward extending our understanding and optimization of the interface between the imaging physics and human image readers to the further interface of these with computerized reading-assist modalities.

APPENDIX

The formulation described in the body of the present article models the situation where the variance structure is allowed to be unequal across modalities by splitting the case, reader, and reader-by-case components in the general linear model, that is, it makes them a function of modality; it leaves the modality-by-case, modality-by-reader, and modality-by-reader-by-case components unsplit. An alternative to this model can be constructed by splitting the latter three components and leaving the former three components unsplit. We present the alternative model equations here, together with a demonstration of a one-to-one correspondence between the two alternative models.

In the alternative model, the modality-by-case, modality-by-reader, and modality-by-case-by-reader terms are functions of modality, *i*, and are written $mc_{ij}(i)$, $mr_{ik}(i)$, and $(mrc)_{ijk}(i)$, respectively. These components are taken to be independent across modalities. The linear model of Equation (1) then becomes

$$\begin{aligned} A_{ijkn} = \mu_i + r_j + c_k + (rc)_{jk} + (mr)_{ij}(i) \\ + (mc)_{ik}(i) + (mrc)_{ijk}(i) + z_{ijkn}(i). \end{aligned} \quad \text{(A1)}$$

The strengths of the components of variance of this model will be distinguished from those of the models discussed in the body of the present article by the addition of a presubscript *C*. Here, as earlier, we consider the case of no replication and thus set ${}_C\sigma^2_\epsilon(i) = {}_C\sigma^2_{mrc}(i) + {}_C\sigma^2_z(i)$. Equations (3), (4), and (7)–(10) for the observable variances in terms of the model components of variance for the case of no replication then become

$$\begin{aligned} \text{var}(A_{RC|M}) = {}_C\sigma^2_r + {}_C\sigma^2_c + {}_C\sigma^2_{mc}(M) \\ + {}_C\sigma^2_{mr}(M) + {}_C\sigma^2_{rc} + {}_C\sigma^2_\epsilon(M), \end{aligned} \quad \text{(A2)}$$

$$\mathrm{var}(A_{C|MR}) = {}_C\sigma_c^2 + {}_C\sigma_{mc}^2(M) + {}_C\sigma_{rc}^2 + {}_C\sigma_\epsilon^2(M), \qquad \text{(A3)}$$

$$\begin{aligned}\mathrm{var}(A_{RC|M} - A_{RC|M'}) = {}& {}_C\sigma_{mr}^2(1) + {}_C\sigma_{mr}^2(2) \\ & + {}_C\sigma_{mc}^2(1) + {}_C\sigma_{mc}^2(2) \\ & + {}_C\sigma_\epsilon^2(1) + {}_C\sigma_\epsilon^2(2), \qquad \text{(A4)}\end{aligned}$$

$$\mathrm{var}(A_{C|RM} - A_{C|R'M}) = 2{}_C\sigma_{rc}^2 + 2{}_C\sigma_\epsilon^2(M), \qquad \text{(A5)}$$

$$\begin{aligned}\mathrm{var}(A_{C|RM} - A_{C|RM'}) = {}& {}_C\sigma_{mc}^2(1) + {}_C\sigma_{mc}^2(2) \\ & + {}_C\sigma_\epsilon^2(1) + {}_C\sigma_\epsilon^2(2), \qquad \text{(A6)}\end{aligned}$$

$$\begin{aligned}\mathrm{var}(A_{C|RM} - A_{C|R'M'}) = {}& 2{}_C\sigma_{rc}^2 + {}_C\sigma_{mc}^2(1) \\ & + {}_C\sigma_{mc}^2(2) + {}_C\sigma_\epsilon^2(1) + {}_C\sigma_\epsilon^2(2), \qquad \text{(A7)}\end{aligned}$$

As with Equations (3), (4), and (7)–(10), these equations also reduce to the expressions in our previous work (4) for the case where the variances are equal across modalities. Notice, however, that Equations (A2)–(A7) are now linear in the model components. Thus, they may be solved for these components by linear algebra in the same manner as was used in our earlier work (4).

As noted earlier, the left-hand sides of Equations (A2)–(A7) are observables that are independent of any model. Thus, we may equate the right-hand sides for the present model with the corresponding right-hand sides of Equations (3), (4), and (7)–(10) to discover the relations between the components of variance in the two models. For example, equating the right-hand sides of Equation (8) and Equation (A5) yields

$${}_B\sigma_{rc}^2(M) + {}_B\sigma_\epsilon^2 = {}_C\sigma_{rc}^2 + {}_C\sigma_\epsilon^2(M), \qquad \text{(A8)}$$

where, as above, the presubscript *B* refers to the model in the body of the present article. Similarly, equating the right-hand sides of Equation (4) and Equation (A3), and subtracting Equation (A8) yields

$${}_B\sigma_c^2(M) + {}_B\sigma_{mc}^2 = {}_C\sigma_c^2 + {}_C\sigma_{mc}^2(M), \qquad \text{(A9)}$$

and equating the right-hand sides of Equation (3) and Equation (A2) and subtracting the results in Equations (A8) and (A9) yields

$${}_B\sigma_r^2(M) + {}_B\sigma_{mr}^2 = {}_C\sigma_r^2 + {}_C\sigma_{mr}^2(M). \qquad \text{(A10)}$$

The complete parallelism of Equations (A8)–(A10) may be more apparent on recalling that σ_ϵ^2 in all models here contains σ_{mrc}^2. These three equations show that changing from model B to model C changes the distribution of variance strength *within* the three compartments defined by these three equations, but it does not redistribute variance strength *across* these three compartments or equations. Continuing in this way, we may solve for the components of model B in terms of those of model C and vice versa, as we now show.

Equating the right-hand sides of Equations (10) and (A7), equating the right-hand sides of Equations (4) and (A3), and subtracting yields

$${}_C\sigma_c^2 = {}_B\sigma_c(1){}_B\sigma_c(2). \qquad \text{(A11)}$$

Finally, the equivalence of the right-hand sides of Equations (9) and (A6), and of Equations (7) and (A4), leads in a similar way to

$${}_C\sigma_{rc}^2 = {}_B\sigma_{rc}(1){}_B\sigma_{rc}(2) \qquad \text{(A12)}$$

and

$${}_C\sigma_r^2 = {}_B\sigma_r(1){}_B\sigma_r(2). \qquad \text{(A13)}$$

Thus, from Equations (A8)–(A10) and Equations (A11)–(A13), we have also

$$\begin{aligned}{}_C\sigma_{mr}^2(M) &= {}_B\sigma_r^2(M) + {}_B\sigma_{mr}^2 - {}_B\sigma_r(1){}_B\sigma_r(2), \\ {}_C\sigma_{mc}^2(M) &= {}_B\sigma_c^2(M) + {}_B\sigma_{mc}^2 - {}_B\sigma_c(1){}_B\sigma_c(2), \\ {}_C\sigma_\epsilon^2(M) &= {}_B\sigma_{rc}^2(M) + {}_B\sigma_\epsilon^2 - {}_B\sigma_{rc}(1){}_B\sigma_{rc}(2). \qquad \text{(A14)}\end{aligned}$$

Equations (A12)–(A14) express the components of model C in terms of the components of model B. The relationships in the other direction may be obtained as follows.

The first equation of the set, Equation (A14), provides two equations whose difference is

$${}_B\sigma_r^2(1) - {}_B\sigma_r^2(2) = {}_C\sigma_{mr}^2(1) - {}_C\sigma_{mr}^2(2). \qquad \text{(A15)}$$

The square of Equation (A13) may be used to rewrite the second term of Equation (A15) in terms of the first (and vice versa), providing a quadratic equation in ${}_B\sigma_r^2(1)$ (or

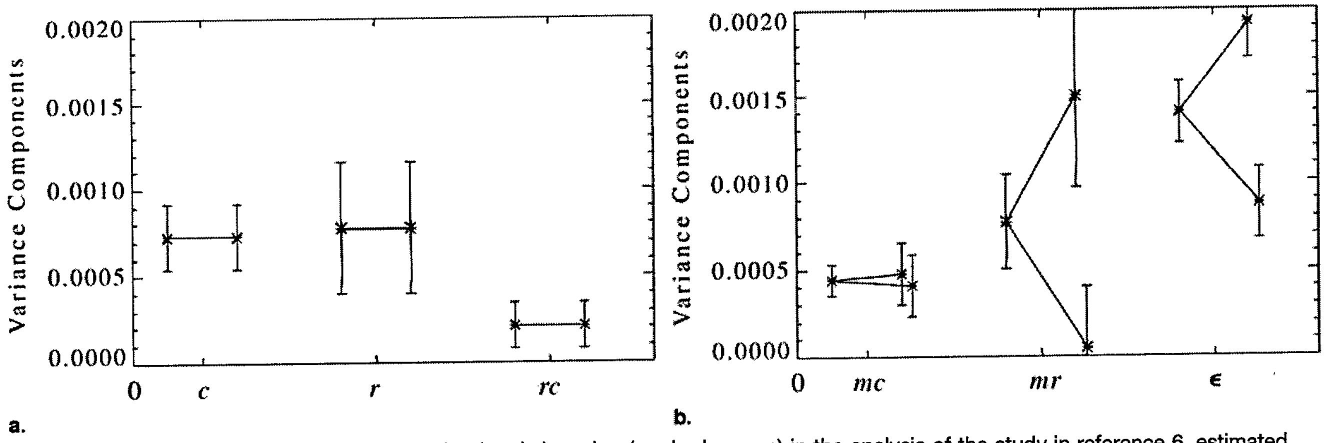

Figure A1. **(a)** Variance components *c* (case), *r* (reader), and *rc* (reader-by-case) in the analysis of the study in reference 6, estimated with model A of reference 4 (vertical bars to the left in each pair) and model C of the Appendix (vertical bars to the right in each pair). Vertical bars represent mean estimates, ± 1 standard deviation obtained with the methods of reference 7. Note that these three components remain unchanged in going from model A to model C. **(b)** Variance components *mc* (modality-by-case), *mr* (modality-by-reader), and ϵ (residual error) in the analysis of the study in reference 6. Vertical bars represent mean estimates, ± 1 standard deviation estimated with the method in reference 7. Unsplit components (to the left in each set of three) are estimated with the model of reference 4 (denoted model A in the present article). Splitting components (the pair to the right in each set) are estimated with model C of the Appendix.

${}_B\sigma^2_r(2)$). The solutions are

$$ {}_B\sigma^2_r(1) = [({}_C\sigma^2_r)^2 + (b/2)^2]^{1/2} + b/2, \qquad (A16)$$

$$ {}_B\sigma^2_r(2) = [({}_C\sigma^2_r)^2 + (b/2)^2]^{1/2} - b/2,$$

where

$$b = {}_C\sigma^2_{mr}(1) - {}_C\sigma^2_{mr}(2).$$

The form of Equation (A16) shows that there is only one nonnegative solution. Parallel solutions of identical form can be found in the same way for ${}_B\sigma^2_c(M)$ and ${}_B\sigma^2_{rc}(M)$.

Finally, expressions for the nonsplitting components in model B may be obtained in terms of the components in model C by combining Equation (A16) and its analogs with Equations (A8)–(A10).

The present exercise demonstrates a one-to-one mapping between model C and model B. The selection between them thus appears to be a matter of intuitive appeal or taste. An appealing feature of model B is that the components that are split correspond to populations (cases, readers, readers-by-cases) that seem intuitively natural, and thus model B is pedagogically attractive. On the other hand, the splitting employed by model C may be more intuitive for some, and an attractive feature of this model is the fact that the equations to which it leads, Equations (A2)–(A7), remain linear in a set of independent variance components. As a consequence, it is suitable for incorporation into conventional ANOVA (as in DBM [2], for example). (The feature of linearity is not an issue for the multiple-bootstrap approach; the choice of model B versus model C leads to only small differences in the computer coding that is required in that approach.) Finally, model C requires no adjustment to accommodate split-plot designs.

In the same manner as above, we may also show that the interaction components in model A are related to those in model C as simple arithmetic averages:

$$ {}_A\sigma^2_{mr} = [{}_C\sigma^2_{mr}(1) + {}_C\sigma^2_{mr}(2)]/2, \qquad (A17)$$

and similarly for σ^2_{mc} and σ^2_ϵ. Now, it is Equation (A4) that determines the confidence intervals on the difference of ROC accuracy measures across two fixed modalities when readers and cases are taken as random effects. The left-hand side of Equation (A4) describes the underlying population or bootstrap experiment. The right-hand side is its decomposition according to model C and is proportional to the sum of three averages, namely, the right-hand side of Equation (A17) and the analogous terms for the σ^2_{mc} and σ^2_ϵ components. The averaged components are precisely the terms that contribute in model A, the equal-variance model. Thus, as far as the confidence interval of interest here is concerned, no new issues arise when moving from model A to model C. (This is the same conclusion found in the body of the article when moving from model A to model B.)

The example of the present article may be analyzed in terms of model C of this Appendix. The results are shown in Figure A1a and A1b. The particular details of these figures are different from those in Figure 1a and 1b of the text, because the absolute levels of the quantities that are split differ across the two models. However, because of the one-to-one correspondence between the two models, there is no fundamental difference between the conclusions drawn from either set of figures.

DEDICATION

The authors dedicate this work to the memory of Donald D. Dorfman, PhD, of the University of Iowa, who passed away on April 15, 2001. Don's singular contributions to this field have always been an inspiration to the present authors. The field will not be the same without him.

REFERENCES

1. Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York, NY: Academic Press, 1982.
2. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27:723–731.
3. Gatsonis CA, Begg CB, Wieand S, eds. Advances in statistical methods for diagnostic radiology: a symposium. Acad Radiol 1995; 2(suppl 1):S1–S84.
4. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative methodology for random-effects ROC analysis. Acad Radiol 2000; 7:341-349.
5. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic index estimates. Acad Radiol 1997; 4:587-600.
6. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 1999; 6:22–33.
7. Beiden SV, Wagner RF, Campbell G, Chan HP. Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis. Acad Radiol 2001; 8:616-622.
8. Dorfman DD, Berbaum KS, Lenth RV, Chen YF. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: split plot experimental design. Proc SPIE 1999; 3663:91-99.
9. Efron B. The jackknife, the bootstrap and other resampling plans. Philadelphia, Pa: Society for Industrial and Applied Mathematics, 1982.
10. Efron B, Tibshirani RJ. An introduction to the bootstrap: monographs on statistics and applied probability. New York, NY: Chapman & Hall, 1993.
11. Goodenough DJ, Metz CE. Implications of a "noisy" observer to data processing techniques. In: Raynaud C, Todd-Pokropek A, eds. Information processing in scintigraphy. Orsay, France: Commissariat a l'Energie Atomique, Departement de Biologie, Service Hospitalier Frederic Joliot, 1975.
12. SAS Institute. SAS/STAT software: changes and enhancements through release 6.12. Cary, NC: SAS Institute, 1997; 573–577.
13. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837–845.
14. Campbell G, Douglas MA, Bailey JJ. Nonparametric comparison of two tests of cardiac function on the same patient population using the entire ROC curve. In: Ripley KL, Murray A, eds. Computers in cardiology. Long Beach, Calif: IEEE Computer Society, 1989; 267–270.
15. Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. The potential of computer-aided diagnosis (CAD) to reduce variability in radiologists' interpretation of mammograms. Radiology (in press).
16. International Commission for Radiation Units and Measurements. Medical imaging: the assessment of image quality. ICRU Report no. 54. Bethesda, Md: International Commission for Radiation Units and Measurements, 1996.
17. Metz CE, Wagner RF, Doi K, Brown DG, Nishikawa RM, Myers KJ. Toward consensus on quantitative assessment of medical imaging systems. Med Phys 1995; 22:1057-1061.
18. Beam C, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med 1996; 156:209-213.

Technology in Cancer Research & Treatment
ISSN 1533-0346
Volume 1, Number 3, June (2002)
©Adenine Press (2002)

Computer-Aided Diagnosis of Breast Cancer in Mammography: Evidence and Potential

Yulei Jiang, Ph.D.

Kurt Rossmann Laboratories for
Radiologic Image Research
Department of Radiology
The University of Chicago
5841 South Maryland Avenue
Chicago, IL 60637

www.tcrt.org

Computer-aided diagnosis (CAD) methods are being developed to help radiologists improve the interpretation of mammograms for the detection of breast cancer. We review several laboratory observer performance studies of computer-aided diagnosis of malignant and benign breast lesions. These studies show that CAD can improve radiologists' diagnostic performance by increasing the number of their biopsy recommendations for actual malignant lesions while reducing the number of their biopsy recommendations for suspicious but actually benign lesions, and by reducing the variability in their interpretation of mammograms. These results indicate a potential clinical role of CAD in mammography for the detection of breast cancer.

Introduction

Breast cancer is the most frequently diagnosed cancer in women and the second leading cause of cancer mortality in women (1). Screening mammograms are effective in detecting asymptomatic breast cancers that can be treated effectively. However, mammography faces challenges to increase sensitivity further and to reduce the number of false-positive mammograms (2). Computer-aided diagnosis (CAD) is proposed to help improve radiologists' diagnostic performance in radiology in general and in mammography in particular. In CAD, a radiologist interprets mammograms that are also analyzed by a computer that detects potential breast lesions or differentiates breast lesions as malignant or benign. The radiologist's interpretation of the mammograms and his or her diagnosis are enhanced by the computer analysis of the same mammograms.

CAD consists of two essential components: an automated computer technique that analyzes the mammograms, and the consideration of the computer analysis results rendered by a radiologist that effects the radiologist's diagnostic decision making. Both components are important. A high-performance computer technique is the essence of any CAD method, and the effect of this computer technique on radiologists' diagnostic decision-making is equally important. An analogy may be drawn between CAD and the current image-based practice in diagnostic radiology. The computer analysis is analogous to the high-quality images that must be interpreted by highly skilled radiologists whose interpretation of the diagnostic images determines diagnostic accuracy.

CAD can be broadly described as either for the purpose of detection or for the purpose of classification. Computer detection techniques identify potential lesions in mammograms, such as masses and clustered microcalcifications, to help radiologists avoid missing subtle lesions that may be small cancers. Computer classification techniques classify lesions into specific diagnostic cate-

Corresponding Author:
Yulei Jiang, Ph.D.
Email: y-jiang@uchicago.edu

gories, such as malignant versus benign, to help radiologists better analyze the lesion once it is detected and decide on an appropriate approach for management. In this report, we focus on computer-aided classification of breast lesions. The purpose of this report is to review the evidence in the literature supporting a potential clinical role of CAD in the diagnosis of malignant and benign breast lesions.

Computer Classification of Breast Lesions

Several computer techniques that classify breast lesions as malignant or benign have been reported in the literature. These techniques use a common general approach that involves the extraction of lesion features from mammograms and an analysis of the lesion features using a statistical classifier. This is loosely modeled after the interpretation process used by the radiologist (3, 4). Table I lists the image features that we extracted from mammograms of clustered microcalcifications. These image features correlate qualitatively with radiologists' perceptual experience and this correlation serves as a basis for the use of these image features in our computer technique to classify microcalcifications as malignant or benign (5). These image features were analyzed and merged into an estimate of the lesion's likelihood of malignancy in our computer technique with an artificial neural network. Other classifiers such as linear discriminant analysis (LDA) have also been used in other techniques (6, 7).

Evidence of Potential Clinical Benefits

To date, the most direct evaluation of computer-aided diagnosis of malignant and benign breast lesions is made in observer performance studies conducted retrospectively in research laboratories. In these studies, radiologists review a set of mammograms without the computer aid and record their diagnostic performance. They then review a set of mammograms with the computer aid and compare their diagnostic performance to their unaided performance. These studies simulate the clinical use of CAD and provide evidence of the potential effects of CAD on improving radiologists' diagnostic performance. The results of these studies are more convincing than a comparison of the computer performance alone and the performance of radiologists because this latter comparison does not represent how CAD will be used clinically. We review an observer performance study that we have performed in our laboratory and describe the results of a few other studies.

An Observer Performance Study

We performed an observer study to evaluate the effects of CAD in the diagnosis of malignant and benign clustered microcalcifications (8). Clustered microcalcifications lead to the diagnosis of approximately half of breast cancers. We used 104 cases of mammograms from a consecutive biopsy series of clustered microcalcifications. In 56 cases the microcalcifications corresponded to a malignant lesion and in 68 cases the microcalcifications corresponded to a benign lesion. Five attending radiologists and 5 senior radiology residents, none of whom had read the study cases prior to the study, reviewed the mammograms. The attending radiologists read mammograms in their routine clinical practice for an average of 9 years (median 6, range 1-30 years) and for an average of 30% of their clinical work, and they read at least 1,000 cases of mammograms in the preceding year. The residents had limited experience in mammography from their residency training. Results from the attending radiologists and residents were analyzed separately and those results were combined only when the differences were small.

The mammograms interpreted by the radiologists consisted of the original films in the mediolateral oblique and cranial-caudal projections of both breasts and magnification views of the microcalcification cluster in the same projections. The microcalcification cluster was identified on all films by wax pencil marks so that the readers were not asked to detect these lesions. A sophisticated study design, sometimes referred to as a counterbalanced study, was used to minimize potential biases that can arise from radiologists reading the same images twice under the unaided and the computer-aided reading conditions (9, 10). The readers read the entire set of mammograms twice in two reading sessions that were separated by 10-60 days (mean 30, median 35 days). In the first reading session, 5 readers read half of the cases without the computer aid and the other half of the cases with the computer aid. In the second reading session, these

Table I
List of Computer-Extracted Image Features for Classification of Clustered Microcalcifications as Malignant or Benign

No.	Computer-Extracted Image Feature	To Characterize Microcalcification:
1	Area of a cluster (i.e., a group of multiple microcalcifications)	Spatial distribution
2	Circularity of a cluster	Spatial distribution
3	Number of microcalcifications in a cluster	Number
4	Average microcalcification area	Size
5	Average effective microcalcification volume (area times effective thickness, effective thickness is calculated from contrast)	Size
6	Relative standard deviation in effective microcalcification volume	Size uniformity
7	Relative standard deviation in effective microcalcification thickness (contrast)	Size uniformity
8	Shape-irregularity of microcalcifications	Shape

readers read the images in the complimentary reading condition. For the other 5 readers, the sequence of the reading conditions was exactly reversed. In the first reading session, these readers read the first half of the cases with the computer aid and the second half of the cases unaided. In the second reading session, they read the images in the complimentary reading condition.

The radiologists reported two assessments after reading each case in each of the two reading sessions. They reported their confidence that the microcalcifications corresponded to a malignant lesion and these confidence data were used to compute ROC curves. They also reported their recommendation for patient management from the choices of (1) surgical biopsy, (2) alternative tissue sampling, (3) short-term follow-up, and (4) routine follow-up. These lesion management recommendation data were used to calculate the sensitivity, specificity, and positive predictive value. The readers were told at the onset of the study that approximately half of the lesions were malignant and they were provided with example images with the computer aid to familiarize them with the use of the computer aid.

Improvement in Diagnostic Accuracy

We analyzed the results of our observer performance study and characterized the radiologists' diagnostic performance with receiver operating characteristic (ROC) analysis and by calculating the sensitivity, specificity, and positive predictive value associated with the radiologists' lesion management recommendations (11, 12). For the ROC analysis, we used the summary performance indices of area under the ROC curve, A_z (13), and a partial area index that is considered to be more clinically relevant, ${}_{0.90}A_z'$ (14), that represents the area under a portion of the ROC curve with sensitivities above 90%. The ROC analysis and statistical significance test were done with the Dorfman, Berbaum, and Metz (DBM) method (15).

Figure 1 shows the ROC curves of the radiologists as a group in the unaided and the computer-aided readings, and of the computer performance alone. A substantial improvement in the radiologists' diagnostic performance is apparent as a result of the use of the computer aid. The A_z value increased from 0.61 in the unaided reading to 0.75 in the computer-aided reading. This improvement was statistically highly significant ($P < 0.0001$). The partial area index, ${}_{0.90}A_z'$, also increased: from 0.05 in the unaided reading to 0.24 in the computer-aided reading ($P < 0.0001$, Student's t-test for paired data). Note that the computer performance alone was better even than the computer-aided radiologists. The A_z value of the computer performance alone was 0.80. This indicates that although the radiologists achieved substantial improvements in performance, they have not realized the full potential of gains in accuracy that was possible from the use of the computer aid. The ROC curves of the attending radiologists and the residents were similar, and the differences between their average A_z values were less than 0.01.

From an analysis of the radiologists' lesion management recommendation data, we found that the average sensitivity of the radiologists increased from 74% in the unaided reading to 87% in the computer-aided reading ($P = 0.0006$). Simultaneously, their average specificity increased from 32% in the unaided reading to 42% in the computer-aided reading ($P = 0.003$). These increases in sensitivity and specificity resulted in an increase in the positive predictive value from 46% in the unaided reading to 55% in the computer-aided reading. In terms of the number of patients that a radiologist made the correct diagnosis, use of the computer aid helped each radiologist, on average, to recommend biopsy for 6.4 addition cancer cases and for 6.0 fewer benign lesions.

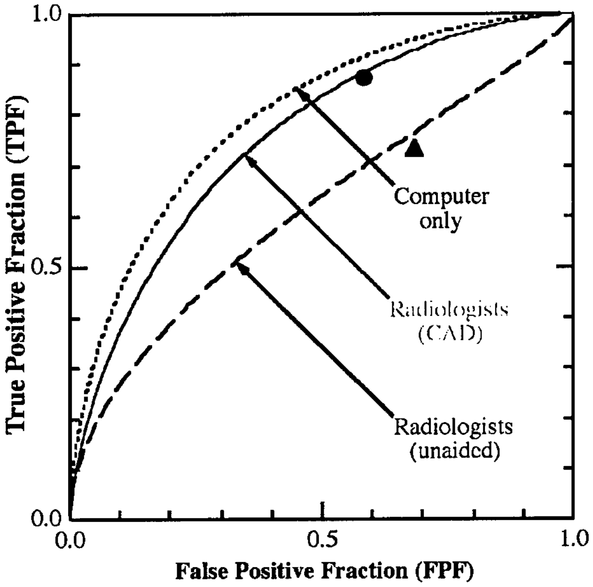

Figure 1: Summary ROC curves of ten radiologists' interpretation of 104 cases of mammograms containing clustered microcalcifications with respect to a malignant or benign diagnosis. The A_z values are 0.61 for the unaided reading and 0.75 for the CAD reading ($P < 0.0001$). As a reference, the computer's A_z value is 0.80. The operating points represent the biopsy recommendations made by the radiologists in the unaided (▲) and the CAD (●) reading conditions.

Reduction of Variability in Mammogram Interpretation

The interpretation of mammograms is influenced by many sources of variations. Radiologists do not always agree with their colleagues in their interpretations of the same mammograms and they do not always agree with themselves in repeated blind interpretations of a single mammogram. Such variability in the mammogram interpretation may be substantial and it lowers the overall effectiveness of mammographic screening for breast cancer (16-18). We analyzed the data of our observer performance study to see if CAD had an effect on reducing the variability of mammogram interpretation (19, 20).

We characterized variability in mammogram interpretation by calculating the agreement and disagreement among the radiologists in their lesion management recommendations, i.e., whether to recommend a biopsy. Agreement was characterized with the kappa statistic (21). Disagreement was characterized with the frequency of substantial disagreement, defined as a situation in which one radiologist recommended biopsy and another radiologist recommended routine follow-up for the same lesion (16). The frequency of substantial disagreement was further calculated in two different ways: in terms of a percentage of the lesions (i.e., patients) in which a substantial disagreement occurred between at least two radiologists who interpreted the patient's mammograms (the per-patient frequency), and in terms of a percentage of all possible recommendation pairs made by any two radiologists who interpreted the same mammogram (the pairwise frequency).

We found that use of the computer aid helped the radiologists to agree more frequently in their lesion management recommendations. The kappa statistic increased from 0.19 (95% CI 0.13 to 0.28) in the unaided reading to 0.41 (95% CI 0.32 to 0.51) in the computer-aided reading. This improvement was statistically significant ($P < 0.05$).

In addition, use of the computer aid also helped the radiologists to reduce the number of substantial disagreements in their lesion management recommendations (Fig. 2). In the unaided reading and calculating only the lesion management recommendations made by the 5 attending radiologists, the pairwise frequency of substantial disagreement was 7% and the per-patient frequency was 23%. Use of the computer aid reduced these frequencies of substantial disagreement among the attending radiologists by 63%. This reduction was statistically significant for all cases combined and for cancer cases ($P < 0.04$), but was not significant for benign cases alone. The residents had more frequent substantial disagreements in their lesion management recommendations. Their pairwise frequency of substantial disagreement was 19% and their per-patient frequency was 51%. Use of the computer aid reduced these frequencies by 28%. This reduction was statistically significant only for all cases combined ($P < 0.04$) and was not significant for the cancer cases or benign cases separately.

In these calculations, the pairwise frequencies are smaller because they represent fractions of 4,680 pairs of recommendations made by two radiologists for the same lesion. In contrast, the per-patient frequencies are larger because they represent fractions of 104 lesions each interpreted by 10 radiologists. A more clinically relevant frequency that corresponds to two radiologists interpreting a patient's mammogram would be between the pairwise and the per-patient frequencies calculated here.

Other CAD Observer Performance Studies

Other mammography observer performance studies have found similar effects of CAD in improving diagnostic performance. Getty *et al.* performed one of the first of such studies (6). In this study, they developed a checklist that a radiologist would fill out as he or she interprets a mammogram. This checklist was intended to guide the radiologist to analyze the appearance of the lesion in a systematic way. By filling out this checklist, the radiologist would also have

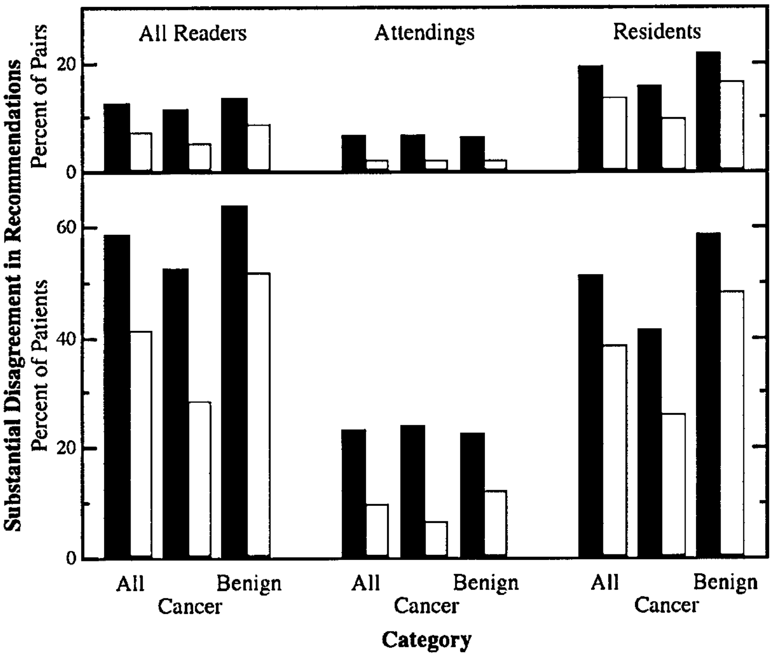

Figure 2. Histograms showing the effect of CAD on substantial disagreement in lesion management recommendations (biopsy vs. routine follow-up). Data shown are pairwise (top) and per-patient (bottom) frequencies. Pairwise frequencies were calculated from all pairs of recommendations made by two different radiologists. Per-patient frequencies were calculated from the total number of cases (104) in which the recommendations were made by multiple radiologists (n = 5 for attending radiologists, n = 5 for residents, and n = 10 for all readers). Black bars = unaided reading and white bars = computer-aided reading. (Reprinted with permission from (19).)

extracted lesion features from the mammogram. These lesion features were then analyzed and merged into an estimate of the probability of malignancy with an LDA classifier. They evaluated the combination of the checklist and the LDA classifier by comparing the reading performance of 6 general radiologists who read 118 cases of mammograms containing breast lesions with and without the aid of the checklist and the LDA classifier (58 lesions were malignant). Both breast masses and clustered microcalcifications were included in the study. The average A_z value of the 6 general radiologists reading the mammograms without the aid of the checklist and the LDA classifier was 0.83. Their average A_z value increased to 0.88 when their mammogram reading was enhanced by the checklist and the LDA classifier. This improvement in diagnostic performance was statistically significant. In addition, they also measured the diagnostic performance of 5 mammography specialists reading the same mammograms without the CAD enhancement. The specialists achieved an average A_z value of 0.88. Therefore, they concluded that the computer aid helped general radiologists to perform at the level of mammography specialists. The disadvantage of this approach, however, is that it requires the radiologist to rate the lesion features interactively as input to the computer classifier.

Chan *et al.* developed a computer technique that classifies breast masses as malignant or benign and evaluated this technique in an observer performance study (7). They used a computer to extract texture-related image features of breast masses and an LDA classifier to calculate a relative malignancy rating of the lesion. Their observer performance study compared the ROC curves of 6 radiologists approved by the Mammography Quality Standard Act who reviewed 238 single-view mammograms with and without their computer aid. They found that use of the computer aid improved the radiologists' diagnostic performance. Their average A_z value increased from 0.87 to 0.91 in the single-view interpretation of the mammograms, and from 0.92 to 0.96 in the interpretation of two-view mammograms. These improvements were statistically significant. Huo *et al.* evaluated in an observer performance study a different computer technique that classifies breast masses as malignant or benign (22). Their computer technique employed computer-extracted image features of lesion morphology and an artificial neural network. They compared in the observer study the performance of 12 radiologists with and without the computer aid and found significant improvement in performance when the computer aid was used (the average A_z increased from 0.93 to 0.96).

Discussion

Results of these observer performance studies show clearly and consistently that radiologists can improve their diagnosis of malignant and benign breast lesions by using computer-aided diagnosis. By considering the computer-estimated likelihood of malignancy of the breast lesion, in a way similar to consulting a fellow radiologist for a second opinion, the radiologist can potentially recommend more malignant lesions to biopsy while recommending fewer benign lesions to biopsy. The radiologist will operate on a higher ROC curve, and achieve greater diagnostic accuracy. In addition to this improvement in diagnostic performance, the radiologist will potentially be more consistent in his or her diagnostic performance, with himself or herself over time, and with the performance of other fellow radiologists. The reduction of variability in radiologists' interpretation of mammograms will be an important added benefit to the improvement of the ROC curves. Both of these improvements can be achieved without having the patient going through an additional examination of another imaging modality and without additional radiation exposure to the patient; these improvements can be achieved from a better use of the information already recorded in a mammogram.

We have focused our discussion of CAD in this report on the diagnosis of malignant and benign breast lesions. The clinical potential of CAD in mammography is not limited to this particular application. CAD methods have been developed to help radiologists detect subtle and early stage breast cancers. Observer performance studies similar to those described in this report have been conducted and have demonstrated that these methods can help radiologists avoid missing subtle breast lesions (23, 24). Some of these methods are now available commercially and are being evaluated in clinical use (25). In addition, methods of computer analysis of mammographic breast density are being developed for analysis of the risk of developing breast cancer (26, 27). These methods promise to identify women at higher risk for developing breast cancer so that early cancer detection may be achieved through better surveillance.

While much research has been done, CAD is still a relatively new concept clinically and it has just begun to enter clinical mammography practice. Over time, one expects to witness the gains in diagnostic accuracy as indicated by the laboratory observer performance studies. The continuous improvement of current CAD techniques and the development of new computer techniques that target other aspects of breast imaging should also increase this gain in diagnostic accuracy over time. CAD could prove to be a powerful and indispensable tool for breast imaging where radiologists face high volume and extremely low cancer prevalence in a screening population that requires their constant vigilance, the challenge of detecting small and curable cancers and the challenge of distinguishing between malignant and benign lesions, and the need to merge information from mammogram, ultrasound, magnetic resonance imaging, nuclear medicine imaging, and patient clinical history.

Acknowledgement

This work was supported in part by a US Army Breast Cancer Research Career Development Award (DAMD17-00-1-0197) and a research grant from Cancer Research Foundation of America.

References and Footnotes

1. Jemal, A., Thomas, A., Murray, T., Thun, M., Cancer statistics, 2002. *CA Cancer J Clin 52*, 23-47 (2002).
2. Elmore, J. G., Barton, M. B., Moceri, V. M., Polk, S., Arena, P. J., Fletcher, S. W., Ten-year risk of false positive screening mammograms and clinical breast examinations. *N Engl J Med 338*, 1089-1096 (1998).
3. Tabár, L., Dean, P. B., Teaching Atlas of Mammography. Thieme-Stratton, New York, (1985).
4. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS™). Third ed. American College of Radiology, Reston, VA, (1998).
5. Jiang, Y., Nishikawa, R. M., Wolverton, D. E., Metz, C. E., Giger, M. L., Schmidt, R. A., Vybomy, C. J., Doi, K., Malignant and benign clustered microcalcifications: automated feature analysis and classification. *Radiology 198*, 671-678 (1996).
6. Getty, D. J., Pickett, R. M., D'Orsi, C. J., Swets, J. A., Enhanced interpretation of diagnostic images. *Invest Radiol 23*, 240-252 (1988).
7. Chan, H. P., Sahiner, B., Helvie, M. A., Petrick, N., Roubidoux, M. A., Wilson, T. E., Adler, D. D., Paramagul, C., Newman, J. S., Sanjay-Gopal, S., Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. *Radiology 212*, 817-827 (1999).
8. Jiang, Y., Nishikawa, R. M., Schmidt, R. A., Metz, C. E., Giger, M. L., Doi, K., Improving breast cancer diagnosis with computer-aided diagnosis. *Acad Radiol 6*, 22-33 (1999).
9. Swets, J. A., Pickett, R. M., Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press, New York, (1982).
10. Metz, C. E., Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol 24*, 234-245 (1989).
11 Metz, C. E., ROC methodology in radiologic imaging. *Invest Radiol 21*, 720-733 (1986).
12. Swets, J. A., Measuring the accuracy of diagnostic systems. *Science 240*, 1285-1293 (1988).
13. Hanley, J. A., Mcneil, B. J., The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*, 29-36 (1982).
14. Jiang, Y., Metz, C. E., Nishikawa, R. M., A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology 201*, 745-750 (1996).
15. Dorfman, D. D., Berbaum, K. S., Metz, C. E., Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Invest Radiol 27*, 723-731 (1992).
16. Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., Feinstein, A. R., Variability in radiologists' interpretations of mammograms. *N Engl J Med 331*, 1493-1499 (1994).
17. Beam, C. A., Layde, P. M., Sullivan, D. C., Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Arch Intern Med 156*, 209-213 (1996).
18. Schmidt, R. A., Newstead, G. M., Linver, M. N., Eklund, G. W., Metz, C. E., Winkler, M. N., Nishikawa, R. M., Mammographic screening sensitivity of general radiologists. In: Digital Mammography. pp. 383-388. Eds., Karssemeijer, N., Thijssen, M., Hendriks, J., Van Erning, L., Kluwer Academic Publishers, Dordrecht (1998).
19. Jiang, Y., Nishikawa, R. M., Schmidt, R. A., Toledano, A. Y., Doi, K., Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. *Radiology 220*, 787-794 (2001).
20. Beiden, S. V., Wagner, R. F., Campbell, G., Metz, C. E., Jiang, Y., Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities. Acad Radiol 8, 605-615 (2001).
21. Cohen, J. A., coefficient of agreement for nominal scales. *Educ Psychol Meas 20*, 37-46 (1960).
22. Huo, Z., Giger, M. L., Vybomy, C. J., Metz, C. E., Effectiveness of CAD in the diagnosis of breast cancer: an observer study on an independent database of mammograms. *Radiology* (in press).
23. Chan, H. P., Doi, K., Vybomy, C. J., Schmidt, R. A., Metz, C. E., Lam, K. L., Ogura, T., Wu, Y. Z., MacMahon, H., Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis. *Invest Radiol 25*, 1102-1110 (1990).
24. Kegelmeyer, W. P., Jr., Pruneda, J. M., Bourland, P. D., Hillis, A., Riggs, M. W., Nipper, M. L., Computer-aided mammographic screening for spiculated lesions. *Radiology 191*, 331-337 (1994).
25. Freer, T. W., Ulissey, M. J., Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology 220*, 781-786 (2001).
26. Boyd, N. F., Lockwood, G. A., Martin, L. J., Knight, J. A., Jong, R. A., Fishell, E., Byng, J. W., Yaffe, M. J., Tritchler, D. L., Mammographic densities and risk of breast cancer among subjects with a family history of this disease. *J Natl Cancer Inst 91*, 1404-1408 (1999).
27. Huo, Z., Giger, M. L., Wolverton, D. E., Zhong, W., Cumming, S., Olopade, O. I., Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: feature selection. *Med Phys 27*, 4-12 (2000).

Date Received: May 5, 2002

Comparison of BI-RADS Lesion Descriptors and Computer-Extracted Image Features for Automated Classification of Malignant and Benign Breast Lesions

Yulei Jiang, Robert M. Nishikawa, Robert A. Schmidt, Carl J. D'Orsi*, Carl J. Vyborny, Maryellen L. Giger, Li Lan, Zhimin Huo, and Alexander V. Edwards

Department of Radiology, The University of Chicago, Chicago, Illinois
*Department of Radiology, Emery University, Atlanta, Georgia
y-jiang@uchicago.edu

Abstract. We compared Breast Imaging Report and Data System (BI-RADS) lesion descriptors provided by radiologists and image features extracted by a computer for computer classification of breast lesions as malignant or benign. Our results indicate that combining the BI-RADS lesion descriptors provided by radiologists and the computer-extracted image features produced the best computer classification performance.

1. Introduction

We are developing computer-aided diagnosis (CAD) techniques to help radiologists more accurately diagnose suspicious malignant and benign breast lesions. Using only computer-extracted image features that we have developed previously, we have shown that our computer techniques can classify breast lesions as malignant or benign as accurately as or more accurately than radiologists [1, 2]. Another approach to computer classification of malignant and benign breast lesions is to use lesion descriptions provided by radiologists in terms of the standard lexicon of the Breast Imaging Report and Data System (BI-RADS) [3, 4]. The purpose of our present study was to compare these two approaches and to investigate whether it is beneficial to combine BI-RADS lesion descriptors provided by radiologists and computer-extracted image features for computer classification of malignant and benign breast lesions.

2. Materials and Methods

The computer-extracted image features of suspicious breast lesions that we have developed previously are described in detail elsewhere. For masses, the computer-extracted image features include two features that characterize spiculations of a mass, margin sharpness of the mass, the density of the mass (average gray level), and texture [5]. For clustered microcalcifications, the computer-extracted image features include the number of microcalcifications within a cluster, the area of the cluster, the circularity of the cluster, the average area of the individual microcalcifications, the average volume of the individual microcalcifications where volume is defined as the product of area and contrast with contrast expressed in terms of a measurement of effective thickness, the relative standard deviation in microcalcification volume, the relative standard deviation in microcalcification effective thickness, and a microcalcification shape irregularity measure [6].

We used the lexicon of lesion descriptions defined in the 3rd edition of BI-RADS (1998) [4]. For masses, the BI-RADS contains mass size (we implemented this descriptor into six categorical measurements from < 1 cm to > 5 cm in 1 cm intervals), 4 descriptors of mass shape, 5 descriptors of mass margin, and 4 descriptors of mass density. Because it is often difficult to select one descriptor to represent accurately the entire margin of a mass, we instructed the radiologists to select up to two margin descriptors for each mass. For clustered microcalcifications, there are 5 descriptors for the distribution of microcalcifications, and 14 descriptors for the morphology of microcalcifications. Similar to the analysis of masses, because it is often difficult to select one morphology descriptor to represent accurately all microcalcifications in a cluster, we instructed the radiologists to select up to two morphological descriptors for each microcalcification cluster. In addition, we added 4 categorical descriptors for the number of microcalcifications in a cluster: <5, 5-10, 10-30, and >30.

We used a database of 92 cases of mass lesions and 127 cases of clustered microcalcifications lesions in this study. The following analysis is on a subset of cases: 67 mass lesions (33 malignant) and 99 microcalcification lesions (42 malignant). In all cases, 4 standard-view mammograms and magnification or spot-compression views of the lesion were available to the radiologists. The standard-view mammograms were digitized to a 100-micron pixel size and analyzed by the computer that extracted the image features. The computer did not analyze the magnification and spot-compression views. Two expert mammographers who are familiar with the BI-RADS standard participated in the study and each of them interpreted all the cases.

The radiologists reviewed the mammograms in a similar fashion as in typical clinical practice where all mammograms of a given case were interpreted together based on which the radiologist reported an overall interpretation. The cases were presented in random order and no time limit was imposed on reading of the mammograms. The radiologists reported their impression either via a laptop computer or verbally to an assistant who recorded the data on the laptop computer. For each case, the radiologists reported on all relevant BI-RADS lesion descriptors for either a mass or a cluster of microcalcifications, the BI-RADS final assessment category, and a quasi-continuous estimate of the likelihood of malignancy.

For the classification of masses, we used a Bayesian artificial neural network [7]. For the classification of clustered microcalcifications, we used the standard feed-forward error back propagation artificial neural network. These classifiers were designed to differentiate between malignant and benign breast lesions. Separate classifiers were designed that used the following as input: (1) the computer-extracted image features only, (2) the BI-RADS lesion descriptors provided by a radiologist only, and (3) the computer-extracted image features *plus* the BI-RADS lesion descriptors provided by a radiologist. In addition, separate classifiers were designed to use as input the BI-RADS lesion descriptors provided by each radiologist. Finally, separate classifiers were designed to classify masses and clustered microcalcifications as malignant or benign. Each classifier was trained and evaluated using the leave-one-out method. Area under the receiver operating characteristic (ROC) curve, A_z, was used as a measure of the performance of the classifiers.

3. Results

Table 1 shows a comparison of the A_z values of the various computer classifiers described above and the A_z values achieved by the radiologists based on their own assessment of the lesions.

Table 1. Summary A_z values of computer classification of malignant and benign breast masses or clustered microcalcifications in comparison to radiologists' assessment of the same lesions.

	Masses		Microcalcifications	
Reader	Reader A	Reader B	Reader A	Reader B
Computer-extracted image features only	0.73	0.73	0.73	0.73
BI-RADS descriptors only	0.88	0.87	0.48	0.72
Computer-extracted image features *plus* BI-RADS descriptors	0.96	0.89	0.75	0.81
BI-RADS final assessment	0.88	0.92	0.60	0.61
Estimate of likelihood of malignancy	0.88	0.91	0.62	0.66

4. Discussion

Computer classification of masses achieved higher A_z values based on the BI-RADS descriptors provided by radiologists than based on the computer-extracted image features. The opposite was true for computer classification of clustered microcalcifications: higher A_z values were obtained based on the computer-extracted image features. The results do not show clearly whether computer classification based on the BI-RADS lesion descriptors provided by radiologists performed better than the radiologists' own assessments in terms of the BI-RADS final assessment categories. In general, the A_z values associated with the classification of masses by either the computer or the radiologists were higher than the corresponding A_z values for the classification of clustered microcalcifications.

The results indicate that combining the computer-extracted image features and the BI-RADS lesion descriptors provided by the radiologists produced the highest A_z values in the classification of both masses and clustered microcalcifications. The combined results were better than results from the computer-extracted image features alone and from the BI-RADS lesion descriptors provided by the radiologists alone.

There is considerable amount of variability in the A_z values based on the BI-RADS lesion descriptors provided by each of the two radiologists. This variability is not unexpected [8, 9]. However, given this variability, we need to collect BI-RADS data from additional expert radiologists and to investigate the effect of this variability on the computer classification of malignant and benign breast lesions.

Acknowledgements

This work was funded in part by NIH through grant R01 CA92361, the US Army Breast Cancer Research Program through grant DAMD17-00-1-0197, and the Illinois Department of Public Health Penny Severns Breast and Cervical Cancer Research Fund. RM Nishikawa, RA Schmidt, CJ Vyborny, ML Giger, and Z Huo are shareholders of R2 Technology Inc., Los Altos, CA. It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interests that would reasonably appear to be directly and significantly affected by the research activities.

References

1. Jiang, Y., Nishikawa, R.M., Schmidt, R.A., Metz, C.E., Giger, M.L. and Doi, K.: Improving Breast Cancer Diagnosis with Computer-Aided Diagnosis. Acad. Radiol. 6 (1999) 22-33
2. Huo, Z., Giger, M.L., Vyborny, C.J. and Metz, C.E.: Effectiveness of CAD in the Diagnosis of Breast Cancer: An Observer Study on an Independent Database of Mammograms. Radiology (in press)
3. Baker, J.A., Kornguth, P.J., Lo, J.Y. and Floyd, C.E.J.: Artificial Neural Network: Improving the Quality of Breast Biopsy Recommendations. Radiology 198 (1996) 131-135
4. American College of Radiology (ACR): Breast Imaging Reporting and Data System (BI-RADS^{tm}). 3rd ed. American College of Radiology, Reston, VA, (1998)
5. Huo, Z., Giger, M.L., Vyborny, C.J., Wolverton, D.E., Schmidt, R.A. and Doi, K.: Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms. Acad. Radiol. 5 (1998) 155-168
6. Jiang, Y., Nishikawa, R.M., Wolverton, D.E., Metz, C.E., Giger, M.L., Schmidt, R.A., Vyborny, C.J. and Doi, K.: Malignant and Benign Clustered Microcalcifications: Automated Feature Analysis and Classification. Radiology 198 (1996) 671-678

7. Kupinski, M.A., Edwards, D.C., Giger, M.L. and Metz, C.E.: Ideal Observer Approximation Using Bayesian Classification Neural Networks. IEEE Trans. Med. Imaging 20 (2001) 886-899
8. Baker, J.A., Kornguth, P.J. and Floyd, C.E., Jr.: Breast Imaging Reporting and Data System Standardized Mammography Lexicon: Observer Variability in Lesion Description. AJR Am. J. Roentgenol. 166 (1996) 773-778
9. Berg, W.A., Campassi, C., Langenberg, P. and Sexton, M.J.: Breast Imaging Reporting and Data System: Inter- and Intraobserver Variability in Feature Analysis and Final Assessment. AJR Am. J. Roentgenol. 174 (2000) 1769-1777

Comparison of Student's t-Test and the Dorfman-Berbaum-Metz (DBM) Method for the Statistical Comparison of Competing Diagnostic Modalities

Yulei Jiang, Department of Radiology, The University of Chicago, Chicago, IL 60637

ABSTRACT

Both Student's t-test for paired data and the Dorfman-Berbaum-Metz (DBM) method report a P value in comparing ROC curves of competing diagnostic modalities. We empirically compared the P values from the t-test and the DBM method using data of two observer studies involving lung-nodule detection (15 readers 240 cases) and breast-lesion classification (10 readers 104 cases). We made 596,637 comparisons based on data drawn from different combinations and subsets of the readers and cases. The average difference in the P values was 0.11 and 0.058 in the lung nodule study (of two separate analyses) and 0.0061 in the breast lesion study. The lung nodule study showed, in the analysis that demonstrated statistical significance with the original full dataset, both $P < 0.05$ or both $P > 0.05$ in 83% of the comparisons. The t-test alone reported $P < 0.05$ in 17%, and the DBM method alone reported $P < 0.05$ in 1% of the comparisons. A second analysis of the part of the lung nodule study that did not show statistical significance with the original full dataset found both $P < 0.05$ or both $P > 0.05$ in 99% of the comparisons. The t-test alone reported $P < 0.05$ in 1%, and the DBM method alone reported $P < 0.05$ in less than 1% of the comparisons. The breast lesion study showed both $P < 0.05$ or both $P > 0.05$ in 91% of the comparisons. The t-test alone reported $P < 0.05$ in 5%, and the DBM method alone reported $P < 0.05$ in 4% of the comparisons. These results indicate that the t-test and the DBM method generally report similar P values, but their conclusions regarding statistical significance often differ and the DBM method should be used because it accounts for both reader and case variances.

Keywords: ROC analysis, statistical test, computer-aided diagnosis

1. INTRODUCTION

Both the Student's t-test[1] for paired data and the Dorfman-Berbaum-Metz (DBM) method of jackknifing and ANOVA[2] can be—and have been—used to compare the diagnostic performance of competing diagnostic modalities. For example, several studies have compared the conventional film interpretation to computer-aided film reading or computer-aided diagnosis (CAD). The area under the ROC curve, A_z[3], is commonly used as a summary index of diagnostic performance, but other indices of diagnostic performance can also be used. In a typical observer study, the A_z value is estimated from a group of multiple readers who interpret a set of multiple cases. Both the t-test and the DBM method report a P value for the differences in the A_z values of the two modalities. In these experiments, because both the cases and the readers are random samples, the total variance of the A_z values will include a case and a reader variance component (plus interactions[2, 4]). Therefore, the P value should account for both the case variance and the reader variance. To apply the Student's t-test, however, one must first reduce the data by calculating an A_z value for each reader and for each modality. By doing so, Student's t-test for paired data cannot account for case variance in its calculation of the P value. Therefore, strictly speaking, the statistical inference of the t-test is valid only for the particular set of study cases[1]. This implies serious limitations for studies that employ Student's t-test as the method of statistical analysis. The DBM method, on the other hand, takes both case variance and reader variance into account in its calculation of the P value and, therefore, its statistical inference is considered valid for the population of cases and the population of readers that are represented by the study[2]. The purpose of the present work is to empirically compare the results of the t-test and the DBM method to substantiate these theoretical considerations.

2. DATASETS

We used two observer study datasets in this empirical comparison of the t-test and the DBM method. The first dataset was that of Freedman et al[5]. This observer study compared radiologists' performance in the detection of solitary lung nodules in plain chest radiographs with and without computer-aided detection. The study consisted of 240 cases (80 abnormal and 160 normal) and 15 radiologist observers. Each observer read all cases three times: first without the computer aid, then after a sufficiently long period of time again without the computer aid and, immediately after reading each case, re-reading of the same case with the computer aid. We call these three reading conditions the independent unaided reading, sequential unaided reading, and computer-aided reading conditions, respectively. The study showed statistically significant improvements in the average A_z values of the radiologists from the two unaided reading conditions to the computer-aided reading condition ($P = 0.0058$ and $P < 0.0001$). In addition, the study did not find the difference in the average A_z values between the independent unaided reading and the sequential unaided reading conditions to be statistically significant ($P = 0.6$).

The second dataset was that of Jiang et al[6]. The purpose of this observer study was to compare radiologists' performance in the diagnosis of malignant and benign clustered microcalcifications in mammograms with and without the aid of a computer technique that provided an estimate of the lesions' likelihood of malignancy. This study consisted of 104 cases (46 malignant and 58 benign) and 10 radiologist observers. Each observer read all cases twice, once without the computer aid and once with the computer aid. A counterbalanced design was used to determine the sequence in which the observers read the mammograms, as described in detail elsewhere[6]. The study found statistically significant improvement in the average A_z values between the unaided reading and the computer-aided reading conditions ($P < 0.0001$).

3. METHODS OF COMPARISON

To compare the t-test and the DBM method, we applied both methods to the same ROC dataset and compared the A_z values and the P values calculated by the two methods. To apply the t-test, we first applied the LABROC4 algorithm[7] to the data of an individual reader to obtain a maximum-likelihood fit of the data to the univariate binormal model and to obtain an estimate of the A_z value. We then used the t-test to evaluate the difference in the means of the A_z values of the two reading conditions using data from a group of readers. The mean A_z values of the two reading conditions and the P value reported by the t-test were recorded. The DBM method was applied as described elsewhere[2]. The ROC data of the group of readers under both reading conditions were used as input to the DBM method simultaneously. The average A_z values of the two reading conditions and the P value associated with the difference in the average A_z values of the two reading conditions were recorded. These recorded A_z values and P values from the t-test and from the DBM method were then compared.

To compare the t-test and the DBM method meaningfully, we applied the two methods on a large number of ROC datasets. These ROC datasets were derived from the two observer study datasets as described below. (1) The two observe study datasets included 15 and 10 readers, respectively. From these full datasets, we obtained data of subsets of readers that consisted of 2-15 readers and 2-10 readers respectively for each study. With a fixed number of readers, different groups of readers can be drawn from the larger pools of readers; we obtained ROC data for each of these groups of readers by performing combination analysis. (2) The two observer study datasets consisted of 240 and 104 cases, respectively. From these full sets of cases, we obtained six subsets of cases from the Freedman et al. study dataset with the total numbers of cases of 177, 178, 185, 198, 222, and 240 in each subset (the last subset was the full case set). The number of normal cases was 160 in all six case subsets. These case subsets were defined in the original study for the purpose of analyzing diagnostic performance of cases of some particular characteristics. But for our present purpose, the details of the characteristics of these case subsets are not important and it suffices to note that the numbers of cases were different among the cases subsets but the individual case in each case subset was held fixed. From the Jiang et al. study dataset, we obtained 201 case subsets. These consisted of 100 randomly selected case subsets of 90 cases (45 malignant, 45 benign), 100 randomly selected case subsets of 50 cases (25 malignant, 25 benign), and the full set of cases of 104 cases (46 malignant, 58 benign). (3) The Freedman et al. study included data of the independent unaided reading and the sequential unaided reading conditions. Comparison of these two reading conditions in the original paper did not show statistical significance. Therefore, this dataset had different statistical conclusion from the other two datasets that compared unaided reading and computer-aided reading conditions. This dataset was therefore included in our present work. The combination of the above three approaches yielded a total of 393,024 ROC datasets from the Freedman et al. study and 203,613 ROC datasets from the Jiang, et al. study. Our comparison of the t-test and the DBM method was then based on all of these ROC datasets.

4. RESULTS

The A_z values calculated by the t-test and by the DBM method were numerically similar. The absolute value of the average difference in the A_z values between that calculated by the t-test and that calculated by the DBM method was 0.0024 in the analysis of the Freedman et al. study dataset that showed statistical significance with the full dataset, 0.0018 in the analysis of the Freedman et al. study dataset that did not show statistical significance with the full dataset, and 0.0033 in the analysis of the Jiang et al. study dataset. Figure 1 shows a histogram of the difference in the A_z values calculated by the two methods.

The P values calculated by the t-test and by the DBM method were similar, on average, but with considerably large differences in some instances. The absolute value of the average difference in the P values calculated by the t-test and by the DBM method was 0.107 in the analysis of the Freedman et al. study dataset that showed statistical significance with the full dataset, 0.058 in the analysis of the Freedman et al. study dataset that did not show statistical significance with the full dataset, and 0.0061 in the analysis of the Jiang et al. study dataset. The median differences were closer to zero in all three analyses, but the range in these differences exceeded 0.2 in both the positive and negative directions. Figure 2 shows a histogram of the differences in the P values calculated by the two methods.

There were both agreements and disagreements between the t-test and the DBM method regarding statistical significance based on the calculated P values. In the analysis of the Freedman et al. study that demonstrated statistical significance

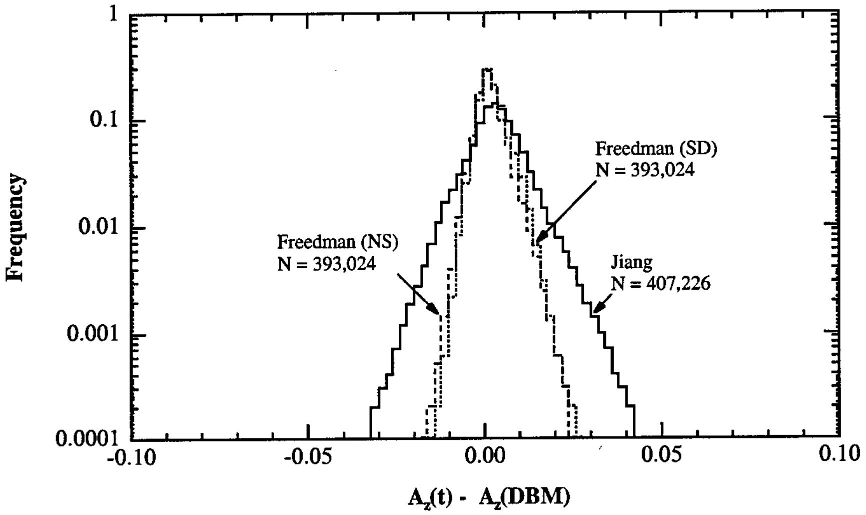

Figure 1. Histograms of the differences between the A_z values computed by the Student's t-test for paired data and by the DBM method. SD = significant difference in the original full dataset. NS = not significant in the original full dataset.

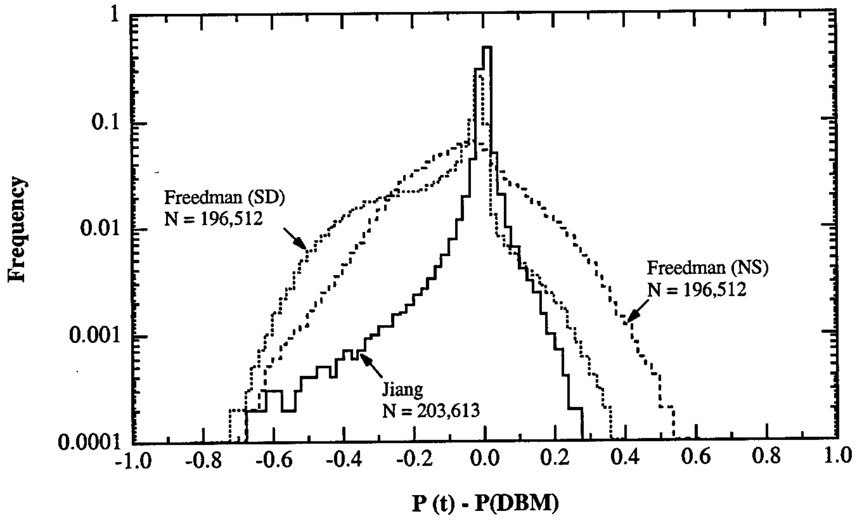

Figure 2. Histograms of the differences between the P values computed by the Student's t-test for paired data and by the DBM method. SD = significant difference in the original full dataset. NS = not significant in the original full dataset.

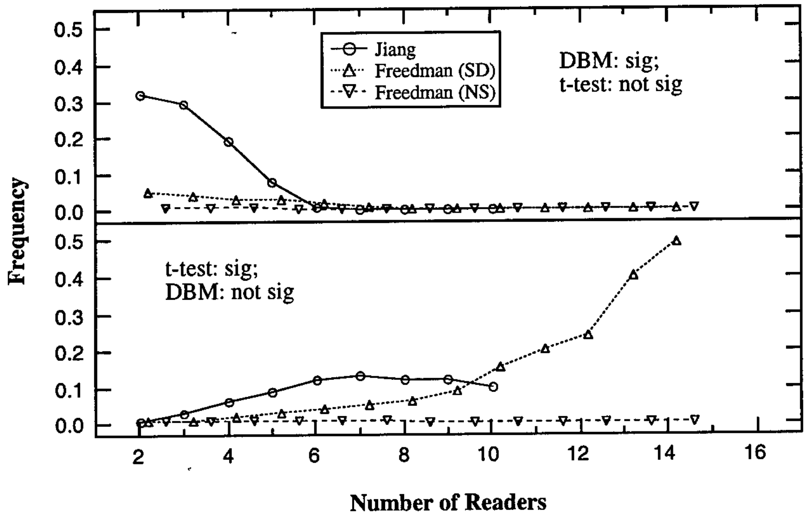

Figure 3. Frequency of disagreement between the t-test and the DBM method in reporting statistically significance at the $\alpha = 0.01$ level as a function of the number of readers. The denominator, which is not a constant, is the number of ROC datasets with a fixed number of readers and with both the individual reader and the number of cases being variable. SD = significant difference in the original full dataset. NS = not significant in the original full dataset.

with the full dataset, both the t-test and the DBM method reported P < 0.05 for 27% of the ROC datasets and neither method reported P < 0.05 for 56% of the ROC datasets. The t-test alone reported P < 0.05 for 17% and the DBM method alone reported P < 0.05 for 1% of the ROC datasets. In the second analysis of the Freedman et al. study that did not show statistical significance with the full dataset, both the t-test and the DBM method reported P < 0.05 for less than 1%, neither method reported P < 0.05 for 99%, the t-test alone reported P < 0.05 for 1%, and the DBM method alone reported P < 0.05 for less than 1% of the ROC datasets. Finally, in the analysis of the Jiang et al. study, both the t-test and the DBM method reported P < 0.05 for 81%, neither method reported P < 0.05 for 10%, the t-test alone reported P < 0.05 for 5%, and the DBM method alone reported P < 0.05 for 4% of the ROC datasets.

The discrepancy between the t-test and the DBM method in statistical significance depended on the number of readers (Fig. 3). With small numbers of readers, the t-test tended to fail to find statistical significance that was found by the DBM method. However, with large numbers of readers, the t-test had a tendency to find statistical significance that was not found by the DBM method.

5. SUMMARY

We compared Student's t-test for paired data and the DBM method in multiple-reader and multiple-case ROC analysis of competing diagnostic modalities using data drawn from two observer studies. A large number of comparisons were made between the t-test and the DBM method. Results show that the two methods report numerically similar A_z values and often report similar P values. However, the two methods also often come to different conclusions of statistical significance depending on the study, the number of readers, etc. Because the DBM method takes both reader variance and case variance into account, and because the t-test takes only reader variance into account, statistical comparisons of diagnostic modalities should be performed using the DBM method (or other similar methods).

6. ACKNOWLEDGEMENTS

We are grateful to the authors of Freedman et al. for use of their observer study dataset in this work, to Dr. Charles Metz for helpful discussions and use of the LABROC4 and LABMRMC algorithms. This work was supported in part by the US Army Medical Research and Materiel Command through grant DAMD17-00-1-0197. Y. Jiang was a consultant to Deus Technology for the observer study data analysis.

REFERENCES

1. C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Invest Radiol* **24**, pp. 234-245, 1989.
2. D. D. Dorfman, K. S. Berbaum and C. E. Metz, "Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method," *Invest Radiol* **27**, pp. 723-731, 1992.
3. J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**, pp. 29-36, 1982.
4. J. A. Swets and R. M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* (Academic Press, New York, 1982).
5. M. Freedman, S. C. B. Lo, F. Lure, X. W. Xu, J. Lin, H. Zhao, T. Osicka and R. Zhang, "Computer aided detection of lung cancer on chest radiographs. Algorithm performance vs. radiologists' performance by size of cancer," *Proc. SPIE* **4319**, pp. 150-159, 2001.
6. Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Acad Radiol* **6**, pp. 22-33, 1999.
7. C. E. Metz, B. A. Herman and J. H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Stat Med* **17**, pp. 1033-1053, 1998.